FUJITSU

White Paper
Free Your Analytics from Paralysis – Infrastructure Matters

Exploring new data sources such as social media, mobile devices and the Internet of Things is a gold mine for business analytics to discover valuable information for business decisions. The dark side: dramatically more data needs to be stored, moved and processed to uncover secrets hidden in the floods of data. Only a powerful and optimized data center infrastructure will enable you to extract, collect and analyze data in the most appropriate way.

Content

I

Entering New Dimensions of Analytics

Apart from human resources, data is the most valuable asset of every organization. Already decades ago, people were aware of this fact and tried to turn their data into value. It was obvious that utilizing this data in an intelligent way for their business could support decisions based on real facts rather than intuition, thus helping improve business processes, minimize risk, reduce costs and increase business in general.

However, they also realized that data in its original form was usually of low value. Therefore data was collected from readily available data sources – mainly transactional databases – and then consolidated and transformed into a form convenient for analysis in order to discover relations, patterns and principles, to finally find the real value. Precisely this was the idea of Business Intelligence (BI) in the early days. The transformed data was loaded and stored in a dedicated database, the so-called Data Warehouse, which was separated from transactional systems in order to unload them from business analytics tasks, reporting or visualization of query results in general. Data Warehouses are optimized for reporting.

Traditional Business Intelligence considers mainly internal and historical views collected from a few data sources. Data is structured and typically stored in a relational database management system. Business analytics tasks are designed against a static data model, and happen periodically - every day, week or month in a batch process. As the average end user isn't trained to do his own sophisticated analysis, the number of direct users initiating queries or dealing with business analytics is strongly limited to a few specialists.

Meanwhile things have changed tremendously, which in turn changes the way of performing data analytics.

There are versatile data sources – internal and external ones – which deserve to be taken into consideration. In addition to transactional databases, it is data from the web, be it blog contents or click streams which can help unveil valuable information, not to forget the content from social media which have evolved to the most commonly used communication platforms. There are multi-media files like video, photo and audio, from which important conclusions for the business can be drawn. There are huge text files including endless logs from IT systems, notes and e-mails which contain indicators that businesses are keen on. And not to forget the vast number of sensors built into things which represent the basis for the Internet of things, which enables mapping the physical world into the virtual world.

It is not just the number of data sources that is increasing; it is also the types of data that proliferate. In the traditional Business Intelligence era, only structured data from relational database management systems was considered. Today, the majority of data – analysts speak of more than 80% - is unstructured.

It is a given, that every sort of data is constantly increasing in terms of volume. It is a thing of the past that data relevant for the business are in the Gigabytes or lower Terabyte range. Today, while being faced with an exponential data explosion, we often speak of hundreds of Terabytes or even Petabytes. Analysts speak of 8 Zettabytes (10^{21} bytes) of data newly generated in 2015, which equals 22 Exabytes (10^{18} bytes) every day. 90% of all data originates from the last 2 years. There is a 65% growth of data per year, which equals a 100% growth every 18 months, or a 12-fold amount of data in 5 years compared to today. Consequently we are talking about Exabytes, Zettabytes and even Yottabytes pretty soon, and we do not recognize any real boundaries.

In contrast to traditional Business Intelligence, where reports were generated by batch processing within hours or days, ad-hoc queries with analytics results in real-time are now demanded in order to take immediate decisions proactively, or even initiate actions automatically. Moreover, the focus of analysis is on prediction about what will happen in the future rather than describing things which happened in the past.

With all this, we have set the stage for one of the megatrends in today's IT and a key element of the digital transformation: This is Big Data.

Big Data combines all the characteristics of data that we have just discussed. Big Data is defined by large data volumes, being of various data types from versatile data sources. Quite often, data is generated at high velocity and needs to be processed and analyzed in real-time. Sometimes data expires at the same high velocity as it is generated. By correlating data and by new ways of analytics, you will generate a new value for your organization, while making many dreams come true. You will be able to better predict what will happen in terms of trends, customer behavior, or business opportunities; you will take better and faster decisions, recognize hidden secrets, skip useless activities and minimize risks. Or in general: you will know things that you have not known before.

Big Data is the New Oil

Big Data is also denoted as the new oil, because there are strong analogies with oil production. Looking at oil production, in a first step, you need to discover the oil springs, from which you will extract the raw oil. Then you will refine the raw oil, thus providing the real value.

With Big Data it is quite similar. At first you need to discover where your treasures are, meaning which data sources you will tap and which data from these are relevant for you. Then you will extract and collect this data. As this data is not in the shape and quality you really need it, you will transform it into a shape which is more appropriate for analytics. Then you will analyze the transformed data and visualize the results, which will hopefully be a sound foundation for taking fast and right decisions and the respective actions.

In both cases, for oil production and for Big Data, two important things are needed to make it happen: this is an infrastructure and services.

How to Get from Big Data to Big Value?

Going for Big Data means have high expectations with regard to valuable results. But having just got much data does not necessarily provide advantages for your business. To get a big value from Big Data, some aspects and questions have to be thoroughly investigated, which will be described subsequently.

Select a use case, in line with your business priorities. It is advisable not to start with several use cases at the same time. Define exactly, what you intend to achieve. Identify the data sources you will tap, as well as the data you need to extract from these sources. If external data sources are involved, you should be sure that they are trustworthy. Figure out, if they are always available when you need them. Define an alternative plan what to do, if they are not available. Moreover, maybe they are available free of charge today. But what if you are charged tomorrow? Are you willing to pay for them, or will you look for an alternative?

Transforming collected data into high quality information is a key requirement. Applying analytics to poor quality data will result in poor output and poor user experience, and the Big Data undertaking will be a waste of time and money. This leads to two important questions which need to be answered: How to transform data into high quality, and how to explore data and discover its meaning?

You should also be clear about what you are looking for in the floods of data. How does the needle in the haystack you want to find look like? Which questions to ask in order to find the needle? And what to do with needle after you have once found it, meaning which actions to trigger. And don't forget to verify, if these actions are feasible and executable.

Which analytic methods are recommendable for your given use case? How visualize the results of your analytics effectively in order to recognize better and faster what you are looking for? Which analytic tools to use and how to use them? How do analytics fit into your processes, and how fast do you need the results? And, as always when dealing with data, you have to ensure security and privacy of Big Data; not always an easy task to be done.

It is worth mentioning that Big Data is not just about hoarding data forever. Think about how long it makes sense to retain data and when to delete it again. Without a proper data lifecycle management, you can easily get lost in the floods of data. The most common obstacle we observe in organizations is that there is too much data and too few resources, as well as a lack of analytic and technical skills. The ideal candidate needed for Big Data has creative ideas, a deep knowledge of data, analytic tools and the intended results.

Infrastructure Matters

Dealing with the aforementioned aspects and questions is for sure a mandatory task in the early phases of any Big Data journey. But once you have coped with all this, the right IT infrastructure in place matters. Without the right infrastructure, you will never be able to free your analytics from paralysis.

Before having a deeper look at the infrastructure options, let us consider the challenges you have to meet. As said, Big Data is about large data volumes today, however tomorrow these volumes will be larger, because we are faced with an exponential growth of data. As the exponential data growth should not have a negative impact on the processing times, the clear objective is to keep processing time constant while data volumes increase. But this is not a target customers want to achieve at any cost; storing and processing these large volumes of data is intended at affordable costs.

Big Data is also characterized by velocity. New tasks come up, queries need to be submitted ad-hoc, and the analytic results are expected in real-time. Finally, there are the event streams continuously generated, for instance by sensors or click streams, which need to be captured and analyzed in real-time in order to take decisions and actions in real-time, too. Depending on the use case, we speak of multiple millions of events per second and latencies in the milliseconds or even microseconds range.

These challenges make it quite obvious that the traditional infrastructure concepts don't really help here. The distinct objectives make it also obvious that one size does not fit all, meaning that various concepts will be part of the play. All told, Big Data changes the infrastructure conversation. Our objective in this chapter is to fill an infrastructure master plan for Big Data step by step, considering all the challenges discussed before. In a first step, data is extracted and collected from versatile data sources. These data sources can be transactional systems and data warehouses, text files, e-mails, multimedia files, IT logs, web logs, Internet pages in general which can even be linked, and social media in particular. As data from these sources is waiting for being extracted and collected, it is denoted as data at rest.

Batch Processing Platform

For the extraction and collection of data from these sources, typically a batch processing platform is used, which also looks after cleansing the collected data (e.g. removing redundancies and contradictions) and the transformation of the initial data into higher quality.

Distributed Parallel Processing

In order to cope with large volumes of data, distributed parallel processing is the concept to be deployed. The idea is to distribute data and I/O to the local disks of many nodes in a server cluster, and move the processing tasks to the nodes where the data to be processed resides. In this architecture basically nothing is shared which leads to an almost unlimited scalability. If a certain number of nodes is not sufficient to complete a certain job in a given period of time, you will just add server nodes to achieve the desired processing time.

In order to make the overall configuration fault-tolerant, data can be automatically replicated to multiple nodes. If a server fails, the respective task can be continued on a server where a data replica resides, fully transparently for software development and operation. Data updates need to be considered for all data copies, otherwise the system would be inconsistent.

As the server cluster can be built from industry standard servers without any specific hardware requirements, distributed parallel processing will prove to be extremely cost-effective and therefore affordable.

The de-facto standard for distributed parallel processing is the open source framework Hadoop. Initially, Hadoop was primarily targeted at batch operation, supported by its core components, the Hadoop Distributed File System (HDFS) which serves as a storage basin for large data volumes, and the Hadoop MapReduce parallelization framework. Over time, Hadoop has evolved to a comprehensive eco-system, supporting other infrastructure concepts needed for Big Data, too.

It is true that the results of the batch processing platform can be directly used for reporting or other purposes. But its main use case is pre-processing large data volumes and their transformation into a shape which is more appropriate for analytics.

Fast Response Platform – Distilled Essence of Big Data

In analogy with oil production, we call the consolidated result of this data transformation the distilled essence of Big Data. The distilled essence is usually much smaller than the initial data volumes, but includes all essential information required for the respective for use case. As analytics applied to the distilled essence will accelerate analytics, we denote the respective platform as fast response platform. Which are the options for this fast response platform?

File Systems

In theory, the distilled essence may be stored in the Hadoop Distributed File System or any other file system, but as with any file system it lacks random read and write access. That's why for queries, a database system is certainly a more recommendable approach.

Relational Databases

As long as the size of the distilled essence is manageable, a relational database or an already existing data warehouse could be a good fit, because people are familiar with the structured query language (SQL). But the larger the database becomes, the longer the queries will take. The row-oriented store of a relational database is perfect for OLTP, but less appropriate for OLAP, because much irrelevant data is read, causing high I/O load and long response times. Relational databases with a column-oriented data store help accelerate analytics, broadening the possibilities. However, it is a matter of fact that a relational database always contains structured data while other data types cannot be managed.

NoSQL Databases

This is where NoSQL databases come into play. NoSQL (Not only SQL) databases are especially designed for Big Data and help overcome the limitations of the relational database paradigm. They are not based on a rigid schema, can easily cope with any data type; they allow format changes on the fly without disrupting applications. Nonetheless, the queries are similar to SQL.

NoSQL databases are designed to be distributed across the nodes of a server cluster and for scale-out, allowing basically an almost linear and unlimited scalability. Replication of data to multiple server nodes enables fault-tolerance and an automatic recovery after failure.

As there is a high speed demand regarding data access and data processing, in many NoSQL implementations a caching function is integrated, keeping frequently used data resident in main memory, thus reducing I/O.

In the field of NoSQL databases there are various data models optimized for different problems. Examples are key-value stores, columnar stores, document stores and graph databases. Some NoSQL databases are part of the Hadoop eco-system.

In-Memory Platforms

It is beyond all questions that the access to data on disk, even if SSD are used, can never be as fast as if data is resident in the memory, and hence closer to the applications. That's why for real-time demands the distilled essence of the transformed data is consolidated into a fast responding in-memory platform.

An in-memory platform is a single server or cluster of servers whose main memory is used for fastest data access. In practice, data accesses are accelerated by a factor of 1,000 to 10,000. The analysis of business data can happen in real-time instead of taking hours, days or even weeks. Important decisions can be taken much faster.

For analytics purposes, no disk storage is needed. However, it must not be ignored that data which is only available in the volatile main memory will be lost after server failure, e.g. caused by a power cut. Using a battery backup for the main memory would help only temporarily.
To prevent data losses, the data contents of main memory must be replicated to a persistent storage. Doing so, several solution approaches are imaginable.

A popular option is the continuous replication of memory data to disk. This guarantees an identical status of main memory and disk at any time. Using Solid State Disks (SSD) will of course accelerate all synchronization activities and above all the recovery after system failure.

Alternatively, in order to reduce I/O load, snapshots of the main memory contents may be generated in regular time intervals or when turning the total system off in a controlled manner. However, a system failure could cause the loss of all updates since the last snapshot. By logging the updates, this gap can be closed, too. The system will be able to recover automatically with the latest valid main memory contents from the last snapshot and the change log. As storage for the data copies, snapshots and the change log, local disks of the server nodes involved in the in-memory platform as well as network storage systems are possible.

To increase data availability and achieve an ultra-fast recovery after server failure, main memory contents can be mirrored between server nodes and therefore kept synchronously. However, data replication over the network takes time and therefore undoes a lot of the time savings gained by in-memory computing. Likewise, data replication reduces the totally available net capacity of the main memory. RDD (Resilient Distributed Datasets) is the approach to avoid both drawbacks. Instead of replicating all data across multiple server nodes, it is just the comparably short data lineage describing how data has been generated which is replicated. Lost data can be rebuilt after failure from the lineage and the last persistent rescue point.

Due to steadily decreasing prices for main memory and the increasing performance of network components which contribute to forming memory contents of several servers to a logical unit, in-memory platforms with ever increasing data capacities become an important building block of infrastructures for Big Data.

Event Processing Platform

The focus of the previous sections was on data at rest, which is waiting for being processed. Now we are going to have a look at data in motion, for instance sensors steadily generating event streams at a high velocity. A throughput of thousands or even millions of events per second is quite normal. It is the task of the event processing platform to collect these event streams and analyze them on the fly. The analysis is based on a set of pre-defined rules that include a condition and an action. If the condition (which may be a correlation of logical and temporal aspects) is fulfilled, actions or alerts will be triggered.

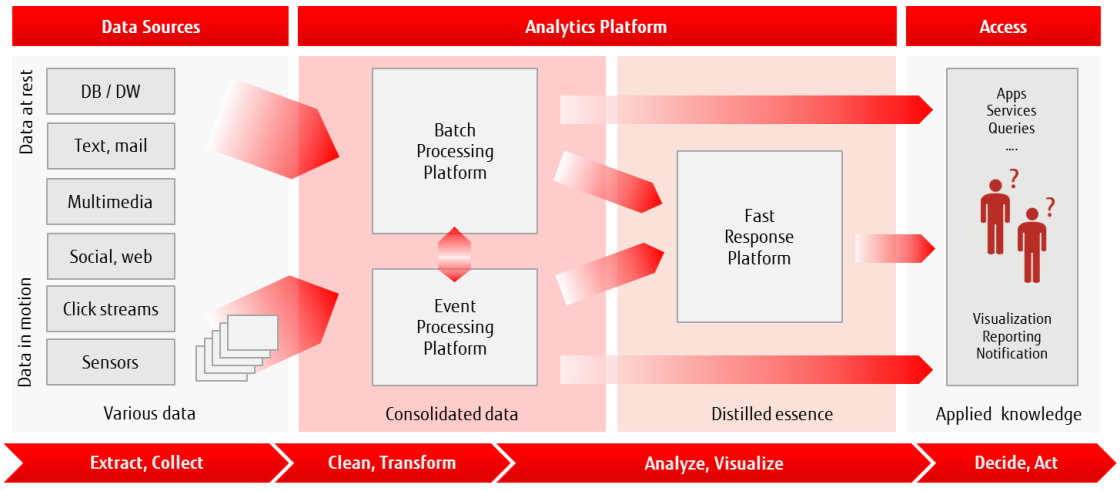
Let us now have a particular look at sensor networks in the context of the Internet of Things. Forwarding the continuously generated event streams immediately through the Internet to the event processing platform in a remote data center will cause huge network traffic and might not be justified, if decisions and actions have only a local effect. The response times over the WAN will be long, and latency demands may perhaps not be met, if they are in the range of milliseconds or even microseconds.

The solution approach is evident. You distribute event processing engines and establish them at the network edge, close to where the event streams are generated. The distributed event processing engines look after data collection and filtering, as well as quick local decisions in real-time. Thus, the traffic through the WAN to the central event processing instance is limited to the filtered data, which are relevant for further analysis and processing at the back-end. This solution approach is also denoted as edge computing or fog computing. In those cases where longer latencies are tolerated, we recommend to do without a de-central engine and use the central one instead, if the network infrastructure is suitable.

The results produced by the central event processing instance may also be forwarded to the batch processing platform or to the fast response platform for further usage.

Big Data Infrastructure Master Plan at a Glance

By now, we have developed the full picture of an infrastructure master plan for Big Data, representing the building blocks from which for every customer-specific situation an optimum solution can be designed by combining them. Depending on the use case, some of the building blocks are only needed alternatively, or they are not even needed at all.



The magnitude of the arrows symbolizes the data volumes transferred between the individual instances of the Big Data infrastructure master plan.

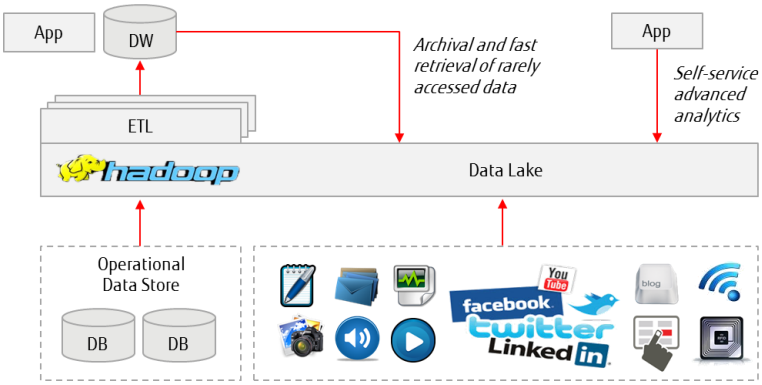
From Data Warehouses to Data Lakes

When starting their Big Data journey, a frequent question raised by organizations is how to proceed with their existing data warehouse. Should they keep it separated from the new Big Data world? Should they combine both worlds? Should they fully replace the data warehouse? What are the reasonable alternatives?

Before answering these questions, let us regard a typical, traditional data warehouse scenario. Data is extracted from the operational data store of transactional systems; it is transformed into a usable shape based on the pre-defined data model, and then loaded into the data warehouse for applying business analytics. This time-consuming process is also denoted as ETL (Extract-Transform-Load).

A data warehouse as it is will not meet your demands forever. Analytics is limited to the pre-defined data model (schema-on-write), and focused on historical things, while descriptive or predictive analytics have their restrictions. New reports, new regulations, new data require a new development which is usually quite expensive. Playing with data, considering data from various angles is impossible due to limited flexibility. As scalability is limited, data volume is also limited. And the cost of data warehouse storage and infrastructure is high. As a matter of fact, data warehouses are often split into departmental siloes, which are fully isolated from each other, preventing data warehouse users to get full picture of all available data. All told, the traditional data warehouse is predominated by limitations.

That's why more and more organizations are going for a radical change. Having understood the advantages of Hadoop, they use a combination of the Hadoop Distributed File System (HDFS) and a NoSQL database (e.g. HBase) as a central data hub for all data coming from the operational data stores. The Hadoop component Hive with an SQL-like interface Hive adds data warehouse functionality to this enterprise data hub.



The idea is to forward only the useful data to the traditional data warehouse by ETL. Moreover, the time-consuming ETL process, which is typically an overnight batch, can be parallelized shortening the ETL time. Due to the horizontal scale properties of Hadoop, one could basically add or reduce commodity server nodes to meet data warehouse reporting service levels as needed.

In a next step, you may add more data sources from which you extract and collect data and store it in its native form in your enterprise data hub, also denoted as the data lake. The schema of this data is only defined when reading it (schema-on-read). Having all data in your data lake means you can apply your analytics on all data. Sensitive data must be encrypted before it enters the data lake.

Your data storage becomes cost-effective and extremely scalable, which even enables you to use the enterprise data hub for archiving data from the data warehouse and fast retrieval of rarely accessed data. Basically, you have the chance to keep all data forever, if it makes sense. Finally, the TCO of your infrastructure will be reduced.

Apart from this, you will become more flexible by using advanced analytics applications you could never apply to the traditional data warehouse model. If these applications can even be used in a self-service like manner by business users, the new world will look perfect. From a logical perspective, this approach is nothing else but an augmentation and an optimization of your data warehouse. And maybe once you will be wondering, if you still need the traditional warehouse. Maybe the Data Lake will offer you already all you need in a flexible manner.

A Data Lake is for sure a primary example of Fast IT. But this concept will work only with well-defined data governance including metadata, measures for security and data quality, master data management and information lifecycle management and the supporting processes.

How to Get to Your Big Data Infrastructure without Paralysis

We have a clear picture now, how a Big Data infrastructure may look like. But how can we get there? There is no doubt that building a data center infrastructure is complex. The main reason for this is the complexity of the infrastructure itself, which is composed of diverse components, such as servers, storage, networks, virtualization layers for all these components, databases and other middleware, as well as applications. In addition, a management layer is needed to keep all components under control.

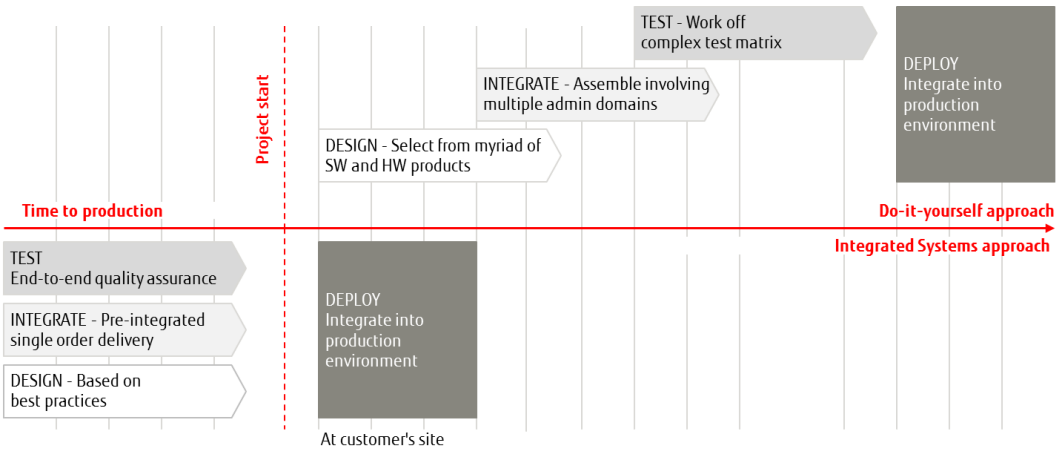
When building the infrastructure in a Do-It-Yourself (DIY) approach, you need to select the right components, sized in the right manner, from a myriad of possibilities, procure and configure them, before you integrate the individual components onsite. As the compatibility of the components is not guaranteed at all, extensive testing is a must. The fact that these components may originate from multiple vendors does not make things easier. All these activities are time-consuming and expensive, while presenting businesses with multiple risks, that things won't work as desired at the end of the day. A deep knowledge of all components involved is required, and an understanding of their various dependencies on each other. Often the coordination among the various administrators who are in charge of the individual components seems to be endless. And as every installation is different, maintenance will be complex, too. Therefore the question suggests itself, if there is a better way to go for Big Data infrastructures.

FUJITSU Integrated System PRIMEFLEX

Of course there is a better way to go: FUJITSU Integrated System PRIMEFLEX, a pre-defined, pre-integrated and pre-tested combination of data center components, such as servers, storage, network connectivity and software with management software being a mandatory part. Based on best practices and real-life project experience, PRIMEFLEX systems are designed in a way that their components will work optimally together.

The benefits resulting from Integrated Systems are manifold. All of a sudden, complexity is reduced. Introducing a new infrastructure in your data center becomes much simpler. You will experience less trouble through trial-and-error testing, because the compatibility of all components is absolutely guaranteed. At the same time, risk is minimized and the skill sets required in your IT department will be less demanding. Apart from this, you need less time for planning, and deployment is tremendously accelerated which shortens time to production and time to value. Due to the optimized design of Integrated Systems, resource utilization is optimized, too. This can have a positive impact on data center space, cabling, energy consumption and cooling efforts. Moreover, an Integrated System represents a perfect foundation for efficient operations and reduced maintenance efforts. All these aspects help reduce cost, both CAPEX and OPEX. Finally, we should not ignore the fact that all these benefits enable IT organizations to focus on the really important aspects of the business. Moving away from a build and maintain focus means improved responsiveness to new business requirements, or even driving business to a new level.

The subsequent figure demonstrates quite impressively the enormous savings in time that can be achieved by choosing the Integrated Systems approach instead of the DIY approach.



With a DIY approach, all typical activities have to be done after the project has started. You have to design the infrastructure, integrate the individual components and test the integrated combination of selected components before the actual onsite deployment and integration into the production environment.

With an Integrated System, necessary things, such as infrastructure design, integration of components and testing, have been done before project start. The required activities after project start are confined to the deployment onsite and the integration into the production environment. Due to all the advantages mentioned, organizations today are adopting Integrated Systems faster than just individual infrastructure building blocks. And this trend will be ever increasing in the future.

What is Specific with PRIMEFLEX?

The PRIMEFLEX line-up includes converged and hyper-converged systems, both built from best-in-class components, either from Fujitsu itself or leading technology partners. PRIMEFLEX systems are either pre-installed in the factory, and arrive ready-to-run at the customer's site; or they are delivered as reference architectures giving you the flexibility to adapt them to your specific requirements. On demand the adjusted configuration can be pre-installed and delivered ready-to-run, combining the advantages of reference architectures with those of ready-to-run systems. For all PRIMEFLEX reference architectures, installation and configuration guidelines are available as a standard.

PRIMEFLEX is supplemented by services throughout all lifecycle phases, either delivered by Fujitsu or our local partners.

Fujitsu has the longest track record in terms of integrated systems; the first system being shipped in 2002. Since then we have continuously optimized or processes for end-to-end solutions, be it in product management, quality assurance, manufacturing and support. Meanwhile we can refer to one of the broadest line-ups in the market and many customer references.

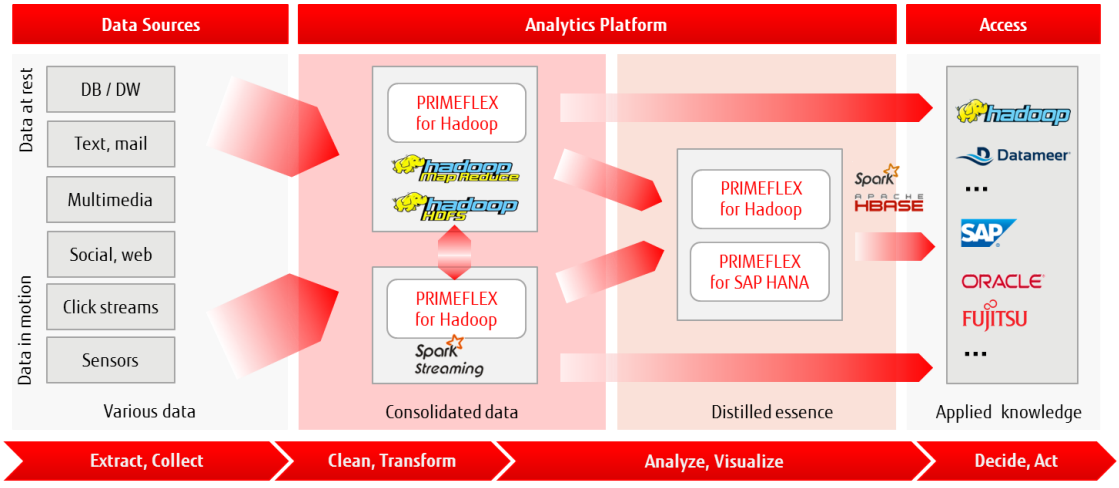
With our PRIMEFLEX family, we address use cases of high relevance. Among these are of course Big Data and Analytics.

PRIMEFLEX meets Big Data

Let us first illustrate which PRIMEFLEX systems are positioned where in our Big Data infrastructure master plan. To make distributed parallel processing easy, we once developed PRIMEFLEX for Hadoop. Due to the evolution of the Hadoop eco-system, PRIMEFLEX for Hadoop can meanwhile be used for running NoSQL databases, and even as an in-memory platform for near-time demands as well as an event processing platform.

Especially for organizations running SAP applications and for those who like the easy to use SAP's integrated development platform for analytics, PRIMEFLEX for HANA is the in-memory platform of choice when it comes to accelerating analytics.

If a relational database management system from Oracle is used as distilled essence, PRIMEFLEX for Oracle Database will be the solution of choice. It is available in a 2-digit number of configuration variants – with and without Oracle RAC (Real Application Cluster), but only available in a limited number of geographies.



PRIMEFLEX for Hadoop

The core building block of Fujitsu's Big Data infrastructures is PRIMEFLEX for Hadoop. It is delivered as a ready-to-run system or reference architecture. There are various configurations which offer choice for our customers in multiple respects. For storage-intensive tasks, we recommend configurations based on PRIMERGY RX rack servers, while for compute-intensive tasks PRIMERGY CX might be the better foundation. A highly sophisticated configuration tool helps identify the optimum cluster size for your needs. PRIMEFLEX for Hadoop can easily scale out for coping with large data volumes, new nodes being integrated into the existing Hadoop cluster in an automated manner. You have also choice in terms of the utilized Hadoop distribution. Starting with Cloudera as a distribution partner, Fujitsu will add further Hadoop distributions on demand.

To relieve Big Data users from developing complex Hadoop MapReduce jobs, PRIMEFLEX for Hadoop includes the optional software component Datameer, which is easy to use for rapid modeling by self-service, even for business users without any deeper IT skills. Thus, Big Data becomes tangible for the business. Datameer covers data collection, analysis and visualization, all in one.

As stated before, PRIMEFLEX for Hadoop can also be used as a platform for NoSQL databases (e.g. HBase), for real-time analytics (based on Apache Spark technology) and for event stream processing (based on Spark Streaming).

PRIMEFLEX for SAP HANA

A commercial in-memory database platform proven for years is SAP HANA. To provide organizations, who have decided for SAP HANA, a fast track to real-time insights, Fujitsu developed the integrated systems PRIMEFLEX for SAP HANA, which is offered as a single node and as multi-node solution in various sizes, ready-to-run or as reference architecture.

The PRIMEFLEX for SAP HANA are either based on FUJITSU Server PRIMERGY or FUJITSU Server PRIMEQUEST. In a single node configuration, it is either the internal disks or a directly attached ETERNUS JX system that serves as persistent storage. With multi-node configurations, an external storage is used, either FUJITSU Storage ETERNUS DX or NetApp FAS. Due to SAP's Tailored Data Center Initiative (TDI), any storage from 3rd parties certified for SAP HANA can be integrated. For development and test purposes, single node configurations can even run in virtual environment based on VMware vSphere.

It is worth mentioning that PRIMEFLEX for SAP HANA includes an SAP certified High Availability and Disaster Recovery configuration option, ensuring highest business continuity.

PRIMEFLEX - More than just Boxes

PRIMEFLEX is not just about hardware and software. It is supplemented by flexible services options throughout the entire lifecycle, delivered by Fujitsu and its local partners. In the consulting phase, it is about identifying the use case to start with, evaluating the achievable benefits, defining the solution architecture and the project roadmap. If Big Data Analytics Services are requested, it will also be about identifying the most appropriate analytics tools and techniques, identifying and preparing the required data, developing analytic models and reports, as well as checking the results.

In the design phase, one of the focus tasks is sizing and the adaptation of the reference architecture in case the available ready-to-run systems do not fit. This is then followed by the on-site deployment and the integration of the new infrastructure into your existing production environment. On demand, additional services, be it for instance the migration to the SAP HANA database, may be necessary.

An integrated system includes various components for which again and again updates and upgrades will be necessary. As the compatibility of all components is not always ensured, high integration and testing efforts have to be taken, the frequency of which can prove as painful due to the different release cycles of the components. Lifecycle management is the pill for headaches caused by updates and upgrades. All relevant updates, patches, security fixes and component replacements are delivered as pre-tested and quality-assured update packages to the customer. This ensures consistency of all components across the entire lifecycle of the Integrated System, which in turn helps reduce maintenance efforts and minimize downtime.

The pill for headaches caused by unpredicted problems during operation is the overall solution support for the entire Integrated System with aligned service levels for all its components, be it hardware or software, be it from Fujitsu or its technology partners. This even applies to reference architectures which have been adapted to customer-specific requirements. It goes without saying that there is a single point of contact for all support matters related to your Integrated System. Beside reactive services based on optimized processes, optional proactive services are offered, comprising a regular system health check. Thus, critical system conditions can be detected early, and preventive maintenance measures will be initiated. Due to Fujitsu's global capabilities, support services can be delivered consistently across geographic borders, or even globally.

If you want to go a step further and disburden your administrators from standard operational tasks, Managed Data Center Services or Managed Hosting Services are worth being discussed. These options represent all-round care-free packages providing peace of mind for IT managers.

And finally, if you have not planned enough budget for your project, we can help out with attractive financing options.

Infrastructure for SQL Databases

If you decide to use a SQL database for your distilled essence, Fujitsu will provide the respective infrastructure, consisting of PRIMERGY or PRIMEQUEST servers, and ETERNUS DX or NetApp FAS storage, supplemented by infrastructure services relieving you from any pain.

High Performance Data Analytics (HPDA)

Analytics is not just about exploring large data volumes of various data types from a diversity of data sources as it was discussed in the context of Big Data. If you consider computer-based modeling and simulations, it is usually structured data from a single source, and it is rather the demand for high computing performance than the size of the data sets which poses a challenge. That's why we speak of High Performance Data Analytics (HPDA) or High Performance Computing (HPC).

High Performance Data Analytics is for instance needed to validate and exploit theories and designs, rapidly build and test virtual prototypes, to increase prototyping coverage and quality, and much more. Respective use cases occur across industries. Examples are crash simulation in the automotive industry, testing new materials in construction and building, drug development against cancer in the pharmaceutical industry, and exploring cosmic origins and the universe, just to mention a few of them.

Working Principle and HPDA Infrastructure

Let us briefly describe the working principle of HPC in general and HPDA in particular: The user defines computing jobs using suitable ISV software and hands the job over to the head node (the management instance) which in turn will distribute the jobs to the compute nodes. Then the execution of the jobs will happen in parallel on the compute nodes which return the results back to the head node, which in turn will return the overall result to the end user.

Typically, the compute nodes and the head nodes are interconnected by a high-speed network, such as InfiniBand. If highest network performance is demanded, it will be helpful to separate the management network from the communication network. For the purpose of high availability it is recommended to use a second head node. The data is stored on shared storage.

Depending on the use case, the number of compute nodes may vary between a few and thousands. The size of the compute node cluster is affected by basically three parameters: the size of the object being the basis for modeling, the accuracy of the model and the expected processing time. The larger the object, the greater the accuracy of the model, and the shorter the demanded processing time, the larger the cluster will be.

HPC can be incredibly complex. It usually means trying to configure, manage, and support your own server cluster – as well as knowing how to write and elaborate scripts. Organizations are understandably nervous about the costs, risks, and potential distractions of stepping into unchartered waters.

PRIMEFLEX for HPC

Here is some good news. It's now possible to harness the power of HPC without all the drawbacks. FUJITSU Integrated System PRIMEFLEX for HPC, based on FUJITSU Server PRIMERGY and FUJITSU Storage ETERNUS DX, lowers the entry barriers to HPC and dramatically simplifies HPC at every step, so you can reap the rewards straight away. Fujitsu's breakthrough in HPC enables even smaller and medium businesses to compete and succeed against their larger counterparts.

The core element of PRIMEFLEX for HPC that brings the HPC cluster to life is Fujitsu's HPC Cluster Suite (HCS), which includes modules for cluster deployment and management, workload management, a parallelization framework, scientific libraries, compilers and other tools as well as the optional Parallel File System FEFS (Fujitsu Exabyte File System). Doubtlessly the HPC Gateway gains the highest attraction among the modules of the HPC Cluster Suite. It provides a desktop-like look & feel, making HPC accessible and usable even by non-IT users. Job submission times are reduced from hours to minutes, multiple applications can be handled easily even if they are distributed over multiple clusters across various locations. Basically, Fujitsu's HPC Gateway embodies the simplification of HPC.

PRIMEFLEX for HPC exists in various flavors. There are ready-to-run systems and reference architectures, both with and without applications pre-installed. Particular appliances are available with application software from ANSYS, COMSOL and Autodesk, especially packaged for smaller and medium-sized organizations. All configurations are Intel Cluster Ready certified.

Analytics and Cloud

There might be many use cases, where you will face with dynamically changing infrastructure demands in your analytics project; be it due to the ever increasing data volumes, be it due to varying performance needs, or be it because the infrastructure is not permanently needed for the same purpose, maybe even alternately for Big Data and HPC. These are exactly the cases where cloud comes into play, making those infrastructure pieces available on demand, which are currently needed. This improves the overall resource utilization, accelerates the delivery of the required infrastructure and reduces the overall cost. Depending on your demands, especially when it comes to security and compliance, public cloud, private cloud or even hybrid cloud will be the solution approach for you.

PRIMEFLEX for Red Hat OpenStack

For those customers that can take advantage from a private cloud infrastructure for Big Data or HPC, PRIMEFLEX for Red Hat OpenStack is the solution of choice. It is about a reference architecture covering the entire cloud infrastructure stack, based on the Red Hat Linux OpenStack platform. Fujitsu's expertise being reflected in the design of PRIMEFLEX for Hadoop, PRIMEFLEX for SAP HANA and PRIMEFLEX for HPC is used in order to deliver the appropriate infrastructure on demand. Moreover, the integrated systems approach leverages Fujitsu's experience in operating OpenStack environments which helps Fujitsu save overall internal IT costs of some 300 Million US dollars within 5 years.

Summary

Big Data and analytics are key prerequisites for the digital transformation. They offer an enormous potential for business value. But they also change the way companies make decisions, do business, succeed or fail. For sure, data and analytic tools play an important part on this journey. But there is no doubt that infrastructure matters. And as so often: one size does not fit all; various concepts should be considered for different use cases. Fujitsu is a one-stop shop for Big Data from which you can always get the optimum infrastructure solution for your use case including software for self-service analytics and all required end-to-end services. With its PRIMEFLEX line-up of Integrated Systems, Fujitsu paves the fast track to data center infrastructures, while minimizing risk and optimizing cost. And if you need choice in terms of sourcing options: with Fujitsu you will get it. All told, digitalizing with Fujitsu equals digitalizing with confidence.

Contact
FUJITSU Technology Solutions GmbH
Address: Mies-van-der-Rohe-Strasse 8,
D-80807 Munich, Germany
Website: www.fujitsu.com/primeflex
2016-05-04 WW EN