Thinking Big Data? Think PaaS

Leveraging the cloud delivery model for Big Data
Introduction

A Nordic survey on Big Data and Hadoop\(^1\) indicated that 92 percent of IT managers of 300 major companies in the Nordics believe that actionable insights acquired from new data analytics would give their company a competitive advantage. More than half (63 percent) saw a need to collect these new types of data, yet many were constrained by traditional databases and systems, ill-equipped to store, process and analyze this data.

Challenging is the increase in this new data, created by incremental digitalization of business processes. For example, in Retail, digitalization has led to new consumer behaviors that have disrupted the industry\(^2\). With this, a new digital footprint has emerged and organizations have a challenge on how to transform it to smart data. However, not only is data growing at an unprecedented rate, creating huge volumes, it’s also streaming in at an unprecedented high speed. To complicate matters, much of it – previously made up of familiar transactional sources – is now unstructured, flowing in from, e.g., social media, sensors, open data, logs, videos and media in various formats.

Companies wishing to exploit the value of Big Data analytics – e.g. acquire better, faster knowledge and intelligence to improve operations and the customer experience – must address the challenges of volume, velocity and variety that this new data poses. According to a 2014 IDC Nordic CXO survey\(^3\), organizations are responding to this data growth through multiple ways. Of these, the four principal approaches are:

- Expand storage capacity
- Upgrade infrastructure
- Re-assess information management
- Implement analytical tools

Indeed, Big Data technologies are transforming the way companies analyze and use data. No longer limited to a sample of subsets, all data – unstructured, semi-structured and structured – can now be analyzed at rest as well as in motion, as it is generated, in real or near-real time.

This is the foundation for creating actionable insights, impacting future organizational strategies. It is business intelligence at its best, looking forward, enabling to ask better questions and predict the future. It is optimizing and fine-tuning an organization’s decision-making and performance through increased automation of business processes, functions, adaptive customer services and business assets.

As companies in the Nordics consider Big Data implementations, they realize such initiatives involve substantial investments in infrastructure, people and competency development. Lack of resources and competency, and uncertainty about the technology are the main obstacles. This white paper discusses a one-stop, comprehensive solution that meets the demands of Big Data analytics. First, the core of a Big Data analytics platform and its three key components are defined. Then the need for integrating with ecosystem components and co-locating platform components are explained. Finally, in contrast to an on-premise infrastructure, consumption of a cloud-delivered platform is proposed for accelerating the adoption of Big Data for Nordic companies, embarking on realizing its benefits.
What is a Big Data platform?

Big Data platforms comprise of technologies that enable acquisition, transformation and the processing of all types of data in a scalable manner that extends beyond the computing limits of individual servers.

In essence, the Big Data platform philosophy is not one of “scale-up” but “scale-out.” In scale-up, scalability challenges are addressed by increasing the computing capacity (CPU or memory), which typically approaches a limit beyond which no additional benefit is realized. In contrast, the scale-out approach uses distributed computing, where data processing is spread across multiple servers, thereby providing near-linear scalability as more nodes are added into the computation cluster.

With large amounts of data processed – ranging from gigabytes to petabytes – coupled with the diversity of the data (structured, semi-structured and unstructured), Big Data solutions address the computation problem by moving the programs closer to the data.

A typical Big Data analytics platform consists of three principal building blocks:

- Big Data distributed ingestion, storage and processing
- Distributed advanced analytics
- Data discovery and visualization

Big Data building block

This building block provides the foundation for the Big Data analytics platform solution. Technically, it is built upon Hadoop (a top-level Apache Foundation project), considered a widely adopted, mature and proven, open-source software framework in the industry. Hadoop also refers to an ecosystem of components that ingest, store and process the data. Furthermore, Hadoop provides various interfaces that allow applications, business intelligence and analytical tools to interact with the data. To fulfill key functions, the Big Data building block should contain the following solutions:

- Data ingestion: Sqoop, Flume, Storm
- Distributed processing and storage: HDFS, Map Reduce, HBase, Spark
- Interfaces: Hive, Impala, Search, Pig, HCatalog, web-HDFS, Mahout
- Workflow management: Oozie
- Data security: Knox, Ranger, Sentry, Navigator

The Big Data building block can be used either by itself or alongside existing massively parallel processing (MPP) data-warehouse solutions.
Advanced analytics building block

The advanced analytics building block has the capability to create sophisticated, statistical models – both descriptive and predictive – that are run against large data sets in a distributed fashion. Advanced analytics commonly covers algorithms that help in predictive or future-oriented analyses.

The advanced analytics building block provides a REST interface, which allows for the execution of algorithms. It also needs to be able to take advantage of the Big Data distributed infrastructure using SQL on Hadoop, and, in future, be compatible with computing frameworks, such as Spark. Rapid model development is enabled by the availability of pre-existing libraries, which should ideally provide:

- Descriptive statistics: covariance, correlation
- Predictive models: linear regression, generalized linear models, logistic regression
- Cluster analysis: K means
- Classification: decision trees, decision forests
- Simulation: Monte Carlo, random numbers

Data discovery building block

This building block provides visualization capabilities, which allow business analysts to explore Big Data visually. It permits the discovery of temporal patterns in data and exploration of relationships among data elements. Furthermore, it provides an omni-channel reporting mechanism, static reports and interactive dashboards. It is also able to perform computations.

The data discovery solution can be connected to multiple data sources – including Hadoop using SQL or SQL like interfaces – while simultaneously...
connected to the REST interface of the advanced analytics building block. The data discovery component may be a commercial, off-the-shelf (OTS) solution or a self-developed application, using visualization libraries, such as d3.js (see illustrations of potential visualization capabilities). The visualization capabilities include Voronoi diagrams, heatmaps, dependency wheels, choropleths, to name a few.

**Co-location of building blocks**

The Big Data platform is intended to efficiently process large amounts of data in the gigabyte to petabyte range. There are, of course, situations where data amounts may not be large, yet still require significant computing power; e.g. in cases of simulations. However, it is highly recommended that all three building blocks are located in the same data center so as to prevent network latency from affecting the solution’s response time.

**Data locality in the Nordics**

A commonly observed pattern in Big Data implementation is the creation of data lakes or data reservoirs. These data bodies collect an organization's digital footprint at a central location. Data assets may include confidential information varying in degree of sensitivity. In such cases, it is vital for companies to ensure that their data is located either in the Nordics or in the European Union. For governments, it is even more essential that the data is located within its own country’s borders. With new legislation on data privacy and security that will be introduced by the European Union through the General Data Protection Regulation (GDPR)\(^4\), data location should be carefully assessed when dealing with personal data. Especially so, when creating Big Data powered solutions to understand customers and markets.

**Integrations**

In addition to these three core building blocks, the platform requires other key solutions with which to integrate – without these, no Big Data solution will be complete. In many cases, companies have already invested into these solutions. If lacking, the service provider needs to provide these through their services portfolio. Components required are:

- Integration platform: to integrate, for example, ERP systems with the analytics engine
- Business rules platform: for automated actions, based on thresholds or scores
- Data management: master data management and data integration tools for managing ETL logic and scripts
Challenges of on-premise infrastructures

Organizations are aware that Big Data is growing at such a fast and furious rate that investment planning is required to handle this data growth – and urgently so as it’s only getting bigger all the time. But solutions that meet Big Data challenges also introduce fresh challenges.

Big Data platforms are commonly architected in the same way as databases: on top of bare-metal (physical) hardware as raw computing power is needed for accessing and processing large quantities of data. The way Hadoop is currently built, makes it assume that data is available locally on each node (server). Similarly, the advanced analytics building block requires very high computation resources, both in terms of memory and CPU.

Another major challenge is operating these large information infrastructures. Distributed computing environments present a far different and greater set of complexities than those of stand-alone, single-server environments. Configuration, administration and maintenance of Big Data platforms also need niche competencies. As such, self-managing such massive infrastructures provides organizations no clear business advantage.

Why is Big Data PaaS a key enabler?

For the purpose of this white paper, a cloud solution refers to the ability of using a platform “as-a-service” consumption-based, delivery model, whereby organizations derive maximum benefit from the OPEX (operational expenditure) advantage. When considering cloud solutions, as with data warehouses, organizations do not require so much elasticity from a cloud solution (being able to scale up or down) as they do scale-out growth, or extensibility – an ability to accommodate an organization’s growth needs. Contrary to an on-premise data center, a cloud solution provides scale-out growth cost-efficiently.

When comparing cloud platforms companies need to make sure that there isn’t an additional, separate network-traffic charge for data ingestion. If so, the enormous amounts of data gathered during this process could incur huge additional charges.

By utilizing a Big Data platform as a service (PaaS), organizations instantly gain:

- The ability to “start small, grow big”
- OPEX benefits
- Focus on Big Data strategy execution

"Vital for companies to ensure that their data is located either in the Nordics or in the European union"

"Another major challenge is operating these large information infrastructures"
Consuming a Big Data Platform as a service (PaaS) further enables companies to focus more clearly on business outcomes and decision-making based on actionable insights.

In addition to improved focus, companies, rather than creating and maintaining one’s own infrastructures, realize clear financial benefits from utilizing a PaaS model. Assuming the same return on investment on a monthly basis based on actionable business insights, the PaaS approach provides a much earlier breakeven than in the case of an on-premise, self-managed platform.

Initial investments for creating a Big Data platform are substantial, with costs of a data center facility, servers, network equipment, personnel and licenses needing an immediate outlay. These upfront investments prevent companies from realizing the value of Big Data in the short term. It is important to note that in the PaaS model, profits realized from actionable insights feed future investment needs to scale out the Big Data platform’s capabilities.

These benefits are key enablers for organizations in the Nordics to incorporate Big Data solutions into their information architecture.
Conclusion

With the increasing digitalization of business processes and new data sources from social media, sensors and e-commerce, organizations have recognized that Big Data presents a new opportunity that is revolutionizing business intelligence, transforming it into a forward-looking process.

At the same time, Big Data is also seen as a challenge, owing to its new complexity as well as its required niche competencies and new infrastructure investments.

By using a cloud-delivery model for a Big Data platform solution, organizations can “begin small and grow big,” without compromising data security, computation needs or cost predictability. Big Data platforms as a service (PaaS) is the key enabler for realizing the value of new Big Data assets.

The changing regulatory landscape around data privacy and data protection makes it necessary to address the issue of data locality when implementing Big Data powered solutions that rely on personal data.

References

1. Big Data Analytics survey conducted by Intel and SAS Institute, Dec 2014
2. Tieto Retail Outlook, Digitaliseringsens framtida Implikationer för detaljhandeln, Feb 2015
3. IDC Nordic CXO Survey, 2014
Tieto is the largest Nordic IT services company providing full life-cycle services for both the private and public sectors and product development services in the field of communications and embedded technologies. The company has global presence through its product development business and global delivery centres. Tieto is committed to developing enterprises and society through IT by realizing new opportunities in customers’ business transformation.

At Tieto, we believe in professional development and results.

Founded 1968, headquartered in Helsinki, Finland and with approximately 14 000 experts, the company operates in over 20 countries with net sales of approximately EUR 1.6 billion. Tieto’s shares are listed on NASDAQ OMX in Helsinki and Stockholm.

Please visit www.tieto.com for more information.