An Outline of Big Data Tools & Technologies

Dinesh Kumar, Assistant Professor, MDU Rohtak, kdinesh@brcm.edu.in

Kamal, Assistant Professor, MDU Rohtak, dkamal@brcm.edu.in

Kanishka Raheja, M.Tech Scholar, MDU Rohtak, kanishka.raheja@gmail.com

1. Overview:

Many of the latest big data technologies are developed by big data startups from around the world that have found a way to deal with the vast amounts of data. They have developed disruptive big data technology that can be used by organizations to obtain valuable insights and turn data into information and wisdom. Of course, also the large existing IT players have developed substantial amounts of big data technology in the recent years. Especially large corporations that want an all-inclusive package installed use those technologies. There are also many different types of analysis that these startups perform and each will have a different impact or result. Some technologies integrate data from different sources directly into a platform, skipping the need for additional data warehousing, while being able to deliver real-time interactive charts that are easy to interact with or to understand. These PaaS (Platform as a service) or DaaS (Data as a Service) solutions allow end-users to work with the data without requiring technical knowledge. There is unanimous agreement that big data is revolutionizing commerce in the 21st century. When it comes to business, big data offers unprecedented insight, improved decision-making and untapped sources of profit. Sometimes, ad hoc tools and applications are the best solution, especially when the objective is to provide information for very specific requirements.

2. Learning from visualizations

There are big data technology vendors that focus on delivering the optimal graphical representation of big data. Visualizing unstructured and structured data is necessary to make the data understandable and turn it into information, but it is also very challenging. New big data startups however seem understand the practice of visualizing and have developed different solutions. One example is visualization based on the visual cortex of the human eye. This maximizes the ability of pattern recognition for the human brain. It makes it easy to read and understand massive amounts of relational data. The use of color and different thicknesses of the threats shown within the cortex allow users to easily recognize patterns and discover abnormalities. Another way of visualizing is to use a technique called topological data analysis. This type of analysis focuses on the shape of complex data and is able to identify clusters and any statistical significance that is present. Data scientist can use this to reveal inherent patterns in those clusters. This type of analyses is best visualized with 3D clusters that show the topological spaces and can be explored interactively. It is definitely not always a necessity to have complex, innovative and interactive graphical representations.

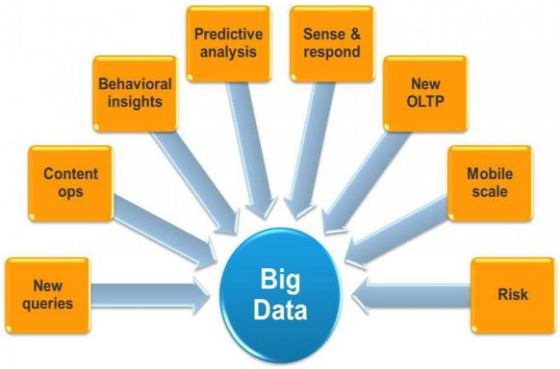

Figure 1: Database scenario for Big Data

Infographics are visual representations of information, data or knowledge and they can help to make difficult and complex material quickly understandable. Dashboards combining different data streams showing 'traditional' graphs (column charts, line charts, pie charts or bar charts) can also provide valuable insights and there are many startups offering such solution. Sometimes, real-time updated simple graphs showing the status of processes already provide more valuable information to improve decision-making then complex innovative visualizations. Visualizations on mobile devices get a completely new meaning when a user is able to play intuitively with the data while swiping, pinching, rotating or zooming on a mobile device.

2.1 Predicting the future

Having real-time analyses visualized in a great way is important, but being able to predict future outcomes will provide even more value to organizations. Analyzing current and historical big data can help to make predictions about future events. This is a huge difference from existing business intelligence, which normally only looks at what has happened using analytical tools but this says nothing about the future. Predictive analysis can help companies provide actionable intelligence based on that same data. Also machine-learning platforms, such as Skytree, can predict trends, make recommendations and reveal untapped markets and customer based on available data. Machine learning goes much further then general business intelligence. Machine learning is about creating algorithms and systems that can learn from the data they process and analyze. The more data processed, the better the algorithm will become.

2.2 Customer Profiling

Profiling of (potential) customers is used to better target customers and better understand (potential) customers. The ultimate goal should be to develop a 360-degrees view of each individual customer, so that eventually an individual offering can be created. Behavioral analytics can be used to discover patterns in (un)structured data across customer touch points, giving organizations better insights in the different types of customers they have. The profiles can also be used within recommendation systems. Of course we have the recommendations from large web shops such as Amazon.com that recommend other products that a user can buy when he or she is in the process of checkout. With big data, real-time recommendations are possible and also more extensive recommendations. Decide for example helps consumers with data-backed recommendations whether to buy now or wait for a new upcoming product. It may be clear that there are so many different possibilities with big data. The global big data landscape with big data startups focusing on different areas is growing rapidly. Therefore we are developing the Big Data Strategy Model to provide some clarification and guidance for organizations in finding the right big data technology. This innovative, one-of-a-kind, model will be revealed soon and will be available for free. It will enable organizations to understand what they can achieve with the data they have, what data they need to develop a certain strategy, which big data technology they need to do that and what big data technology vendor could help them with achieving that strategy.

3. Selecting Big Data Technology: Operational vs. Analytical

The Big Data landscape is dominated by two classes of technology: systems that provide operational capabilities for real-time, interactive workloads where data is primarily captured and stored; and systems that provide analytical capabilities for retrospective, complex analysis that may touch most or all of the data. These classes of technology are complementary and frequently deployed together. Operational and analytical workloads for Big Data present opposing requirements and systems have evolved to address their particular demands separately and in very different ways. Each has driven the creation of new technology architectures. Operational systems, such as the NoSQL databases, focus on servicing highly concurrent requests while exhibiting low latency for responses operating on highly selective access criteria. Analytical systems, on the other hand, tend to focus on high throughput; queries can be very complex and touch most if not all of the data in the system at any time. Both systems tend to operate over many servers operating in a cluster, managing tens or hundreds of terabytes of data across billions of records.

3.1 Operational Big Data

For operational Big Data workloads, NoSQL Big Data systems such as document databases have emerged to address a broad set of applications, and other architectures, such as key-value stores, column family stores, and graph databases are optimized for more specific applications. NoSQL technologies, which were developed to address the shortcomings of relational databases in the modern computing environment, are faster and scale much more quickly and inexpensively than relational databases. Critically, NoSQL Big Data systems are designed to take advantage of new cloud computing architectures that have emerged over the past decade to allow massive computations to be run inexpensively and efficiently. This makes operational Big Data workloads much easier to manage, and cheaper and faster to implement. In addition to user interactions with data, most operational systems need to provide some degree of real-time intelligence about the active data in the system. For example in a multi-user game or financial application, aggregates for user activities or instrument performance are displayed to users to inform their next actions. Some NoSQL systems can provide insights into patterns and trends based on real-time data with minimal coding and without the need for data scientists and additional infrastructure.

	Operational	**Analytical**
Latency	1 ms - 100 ms	1 min - 100 min
Concurrency	1000 - 100,000	1 - 10
Access Pattern	Writes and Reads	Reads
Queries	Selective	Unselective
Data Scope End User	Operational Customer	Retrospective Data Scientist
Technology	NoSQL	MapReduce, MPP Database

Table 1: Overview of Operational vs. Analytical Systems

3.2 Analytical Big Data

These technologies are also a reaction to the limitations of traditional relational databases and their lack of ability to scale beyond the resources of a single server. Furthermore, MapReduce provides a new method of analyzing data that is complementary to the capabilities provided by SQL. As applications gain traction and their users generate increasing volumes of data, there are a number of retrospective analytical workloads that provide real value to the business. Where these workloads involve algorithms that are more sophisticated than simple aggregation, MapReduce has emerged as the first choice for analytics for Big Data. Some NoSQL systems provide native MapReduce functionality that allows for analytics to be performed on operational data in place. Alternately, data can be copied from NoSQL systems into analytical systems such as Hadoop for MapReduce.

3.3 Combining Operational and Analytical Technologies; Using Hadoop

New technologies like NoSQL, MPP databases, and Hadoop have emerged to address Big Data challenges and to enable new types of products and services to be delivered by the business. One of the most common ways companies are leveraging the capabilities of both systems is by integrating a NoSQL database such as MongoDB with Hadoop. The connection is easily made by existing APIs and allows analysts and data scientists to perform complex, retroactive queries for analysis and insights while maintaining the efficiency and ease-of-use of a NoSQL database. NoSQL, MPP databases and Hadoop are complementary: NoSQL systems should be used to capture Big Data and provide operational intelligence to users and MPP databases and Hadoop should be used to provide analytical insight for analysts and data scientists. Together, NoSQL, MPP databases and Hadoop enable businesses to capitalize on Big Data.

4. Which Technologies should we use?

4.1 The user tools: Sometimes, ad hoc tools and applications are the best solution, especially when the objective is to provide information for very specific requirements. For example, the results may be integrated in e-commerce and online services or fundamental for customer and technical support services. In other situations, where reports reflect more typical results in terms of charts, graphs, pivots or matrices of results, standard reporting and BI tools may be the right solution. In this area, a large number of purely commercial products are paired with some open source products that have commercial support. The clear trend is to provide these tools "as a service", so that customers can benefit from a vast scalability for their analysis. The "as a service model" is extremely cost-effective to start a project. One think go consider, however, is if it is possible to change this approach if it becomes less convenient compared to others in the future. This may require the need to re-engineer all the operations.

4.2 The analysis software: Today, the majority of the software used to analyze different information is commercial or open source with commercial support. Software varies in terms of type of data and objectives. For structured data, data mining tools have been around for many years and have reached their maturity so they can be effectively used in Big Data. Map/Reduce operations can crunch a large number of structured data and find patterns or show behaviors that could have not been feasible only few years ago. Extensions to this software can successfully analyze unstructured data, in terms of collected text, documents, images, as well as audio and video streams. These extensions can find similarities in video and audio clips, or in photos. They can understand not only the text stored in a document but also the sentiment and the emotions expressed in collected comments and texts.

4.3 Modeling software: Data modeling is very much related to the kind of analysis to perform. In general, this software comes with the analysis, but some technologies can be alternatives and follow different approaches.

4.4 Infrastructure: This is the area where commodity hardware and open source software is mostly used. There are many commercial solutions that promise improved performance and optimizations, but the common trend is to use less expensive and generally available boxes and products.

4.5 Choice of database: The technology at the center of a Big Data project is, without any doubt, the database. There are several commercial options for Big Data, but the common trend is in the open source area. The set of products developed by the Apache Foundation under the Hadoop umbrella and many side projects are extremely popular and they are considered the de facto standard for Big Data. Truth is, Hadoop can solve only some aspects of the analysis required for Big Data and it may be necessary to pair with other database technologies. NoSQL technologies are particularly popular for their scalability and performance. Cassandra, for example, is a technology mainly used to store a large set of data collection. It is very effective for fast data inserts rather than for analysis. Therefore many projects see Cassandra used to store the acquired data with a denormalised model. MongoDB is used in many cases to store documents or unstructured data in general. Data can be later reviewed and analyzed so that more structured information can be stored in other databases. SQL databases are always a viable choice for Big Data, although they seem to be less popular than Hadoop, Cassandra and MongoDB. Due to their internal architecture, relational databases may struggle if the data acquired is unstructured or it is organized in large objects, such as documents and multimedia clips. In the recent years, much has been done in this area, so relational databases

today are very different from the ones that were used 10 or more years ago. Certainly, the handling and the analysis of structured data is where relational databases can play a leading role. Modern relational databases combine the efficiency of SQL with functionality that can provide faster indexing and optimized access to the data. Columnar relational databases provide for great improvements in traditional data analysis. New indexing algorithms also solve the nuisance of data statistics rebuild, index optimization and storage inefficiency when data is moved in large sets. In addition to these aspects, some relational databases also provide a map/reduce approach similar to the one available in Hadoop and in other NoSQL products. MariaDB is a drop-in replacement for MySQL, the most used open source database for online applications. MariaDB falls into the category of the NewSQL products, i.e. a product that provides unique NoSQL features together with the typical features available in relational databases. Therefore, aspects like transaction management, durability and consistency are available together with schema or schema-less modelling, full text storage and analysis and integration with other NoSQL technologies.

5. Emerging technologies for Big Data

There is an infinite number and combination of different technologies that can be used to give you the perception that the project is already landing somewhere in terms of requirements.

- **Column-oriented databases:** Traditional, row-oriented databases are excellent for online transaction processing with high update speeds, but they fall short on query performance as the data volumes grow and as data become more unstructured. Column-oriented databases store data with a focus on columns, instead of rows, allowing for huge data compression and very fast query times. The downside to these databases is that they will generally only allow batch updates, having a much slower update time than traditional models.
- **Schema-less databases, or NoSQL databases**: There are several database types that fit into this category, such as key-value stores and document stores, which focus on the storage and retrieval of large volumes of unstructured, semi-structured, or even structured data. They achieve performance gains by doing away with some (or all) of the restrictions traditionally associated with conventional databases, such as read-write consistency, in exchange for scalability and distributed processing.
- **MapReduce**: This is a programming paradigm that allows for massive job execution scalability against thousands of servers or clusters of servers. Any MapReduce implementation consists of two tasks. One the "Map" task, where an input dataset is converted into a different set of key/value pairs, or tuples; other the "Reduce" task, where several of the outputs of the "Map" task are combined to form a reduced set of tuples (hence the name).
- **Hadoop**: Hadoop is by far the most popular implementation of MapReduce, being an entirely open source platform for handling Big Data. It is flexible enough to be able to work with multiple data sources, either aggregating multiple sources of data in order to do large scale processing, or even reading data from a database in order to run processor-intensive machine learning jobs. It has several different applications, but one of the top use cases is for large volumes of constantly changing data, such as location-based data from weather or traffic sensors, web-based or social media data, or machine-to-machine transactional data.
- **Hive**: Hive is a "SQL-like" bridge that allows conventional BI applications to run queries against a Hadoop cluster. It was developed originally by Facebook, but has been made open source for some time now, and it's a higher-level abstraction of the Hadoop framework that allows anyone to make queries against data stored in a Hadoop cluster just as if they were manipulating a conventional data store. It amplifies the reach of Hadoop, making it more familiar for BI users.
- **PIG**: PIG is another bridge that tries to bring Hadoop closer to the realities of developers and business users, similar to Hive. Unlike Hive, however, PIG consists of a "Perl-like" language that allows for query execution over data stored on a Hadoop cluster, instead of a "SQL-like" language. PIG was developed by Yahoo!, and, just like Hive, has also been made fully open source.
- **WibiData:** WibiData is a combination of web analytics with Hadoop, being built on top of HBase, which is itself a database layer on top of Hadoop. It allows web sites to better explore and work with their user data, enabling real-time responses to user behavior, such as serving personalized content, recommendations and decisions.
- **PLATFORA:** Perhaps the greatest limitation of Hadoop is that it is a very low-level implementation of MapReduce, requiring extensive developer knowledge to operate. Between preparing, testing and running jobs, a full cycle can take hours, eliminating the interactivity that users enjoyed with conventional

databases. PLATFORA is a platform that turns user's queries into Hadoop jobs automatically, thus creating an abstraction layer that anyone can exploit to simplify and organize datasets stored in Hadoop.

- **SkyTree:** SkyTree is a high-performance machine learning and data analytics platform focused specifically on handling Big Data. Machine learning, in turn, is an essential part of Big Data, since the massive data volumes make manual exploration, or even conventional automated exploration methods unfeasible or too expensive.

6. Conclusions: When it comes to Big Data, it is not always necessary to move away from well-known technologies like relational databases. Modern NewSQL databases like MariaDB can achieve the objective and provide all the features required. The result is a smoother learning curve, less risk, reuse of known technologies and resources and ultimately a reduced total cost of a Big Data project. In other cases, MariaDB can be used in conjunction with NoSQL technologies and integrated in many different ways. To that end, the most important point to consider is that when a single technology is not enough for a successful project, ease of integration is a must.

References:

1. White, Tom (10 May 2012). *Hadoop: The Definitive Guide*. O'Reilly Media. p. 3. ISBN 978-1-4493-3877-0.
2. "Data, data everywhere". *The Economist*. 25 February 2010. Retrieved 9 December 2012.
3. "IBM What is big data? — Bringing big data to the enterprise". www.ibm.com. Retrieved 2013-08-26.
4. "Big Data Definition". MIKE2.0. Retrieved 9 March 2013.
5. Bernhard Warner (April 25, 2013). "'Big Data' Researchers Turn to Google to Beat the Markets". *Bloomberg Businessweek*. Retrieved August 9, 2013.
6. Kalil, Tom. "Big Data is a Big Deal". White House. Retrieved 26 September 2012.
7. Kusnetzky, Dan. "What is "Big Data?"". ZDNet.
8. Hellerstein, Joe (9 November 2008). "Parallel Programming in the Age of Big Data". *Gigaom Blog*.