BIG DATA: TECHNICAL CHALLENGES TOWARDS THE FUTURE AND ITS EMERGING TRENDS

Renuka Sabapathi[1]
Sony Yadav[2]

Abstract

This paper exhibits information related to technical challenges and latest emerging trends relate to big data.Big data is a popular term used to describe the exponential growth and availability of data, both structured and unstructured. The data flow so fast that the total accumulation of the past two years—a zettabyte—dwarfs the prior record of human civilization. But it is not the quantity of data that is revolutionary. "The big data revolution is that now we can do something with the data. Big data is changing the way people within organizations work together. Insights from big data can enable all employees to make better decisions—optimizing operations, preventing threats and fraud, and capitalizing on new sources of revenue.

In the spirit of capturing and describing this big data, this paper highlights the most noteworthy emerging trends such as **Column-oriented databases, Schema-less databases, or NoSQL databases, MapReduce, Hadoop, Hive, WibiData, PLATFORA ,** Big Data in the cloud.

Keywords: Big data, hadoop, HDFS, Nosql, mapReduce

Introduction

Big data is data that is too large to process using traditional methods. It originated with Web search companies who had the problem of querying very large distributed aggregations of loosely-structured data. Big data is data that exceeds the processing capacity of conventional database systems. Big data has become viable as cost-effective approaches have emerged to tame the volume, velocity and variability of massive data. Within this data lie valuable patterns and information, previously hidden because of the amount of work required to extract them.

Characteristics Of Big Data

Very large, distributed aggregations of loosely structured data – often incomplete and inaccessible:

- Petabytes/exabytes of data,
- Millions/billions of people,
- Billions/trillions of records,
- Loosely-structured and often distributed data,
- Flat schemas with few complex interrelationships,

[1] Asst.Prof., B.N.Bandodkar College of Science, Thane, renuka.prema@gmail.com

[2] Asst.Prof., B.N.Bandodkar College of Science, Thane, soni.ryadav16@gmail.com

- Often involving time-stamped events,
- Often made up of incomplete data,
- Often including connections between data elements that must be probabilistically inferred,

Applications that involved Big-data can be:

- Transactional (e.g., Facebook, PhotoBox), or,
- Analytic (e.g., ClickFox, Merced Applications).

Components Of Big Data Processing:

Big-data projects have a number of different layers of abstraction from abstraction of the data through running analytics against the abstracted data. Hadoop is often at the center of Big-data projects, but it is not a prerequisite.

The components of analytical Big-data

1. Packaging and support of Hadoop by organizations such as Cloudera; to include Map Reduce - essentially the compute layer of big data.
2. File-Systems such as the Hadoop Distributed File System (HDFS), which manages the retrieval and storing of data and metadata required for computation. Other file systems or databases such as Hbase (a NoSQL tabular store) or Cassandra (a NoSQL Eventually-consistent key-value store) can also be used.
3. Instead of writing in JAVA, higher level languages as Pig (part of Hadoop) can be used such, simplifying the writing of computations.
4. Hive is a Data Warehouse layer built on top of Hadoop, developed by Facebook programmers.
5.Cascading is a thin Java library that sits on top of Hadoop that allows suites of Map Reduce jobs to be run and managed as a unit. It is widely used to develop special tools.
6.Semi-automated modeling tools such as CR-X allow models to develop interactively at great speed, and can help set up the database that will run the analytics.
7. Specialized scale-out analytic databases such as Greenplum or Netezza with very fast loading load & reload the data for the analytic models
8. ISV big data analytical packages such as ClickFox and Merced run against the database to help address the business issues (e.g., the customer satisfaction issues mentioned in the introduction).

Big Data Infrastructure:

- **Amazon Web Service** a collection of remote computing services, also called web services, make up a computing platform offered by Amazon.com.
- **Cloudera** is the leader in next generation data management. In addition, Cloudera is the leading innovator in and largest contributor to the open source Apache Hadoop ecosystem.
- **BlueData Software** platform makes it easier, faster, and more cost-effective to deploy Hadoop and Big Data infrastructure on-premises.

- **Horton works** Data Platform (HDP), is an enterprise-grade data management platform that enables a centralized architecture for running batch, interactive and real-time applications simultaneously across a shared dataset.
- **IBM** big data solutions can capture manage and analyze huge volumes of structured and unstructured data to improve business insights.
- **MapR** Apache Hadoop distribution claims to provide full data protection, no single points of failure, improved performance, and dramatic ease of use advantages.
- **Snowflake Computing** can safely store, transform and analyze business data, making it easy for everyone to quickly gain insight.
- **Syncsort** provides enterprise software that allows organizations to collect, integrate, sort and distribute more data in less time, with fewer resources and lower costs.
- **Teradata** is engineered for all new data types, offering integrated analytics and revolutionary ways of analyzing data.
- **Treasure Data** helps to collect, analyze, and act on the data safely and efficiently.

Open Source Tool

- **Apache Avro** is a data serialization system.
- **Apache Chukwa** is an open source data collection system for monitoring large distributed systems.
- **Apache Flume** is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data.
- **Apache Hadoop** project develops open-source software for reliable, scalable, distributed computing.
- **HPCC System** is a massive parallel-processing computing platform that solves Big Data problems. The platform is Open Source!
- **Apache Lucene** is a high-performance, full-featured text search engine library written entirely in Java.
- **Apache Oozie** is a workflow scheduler system to manage Apache Hadoop jobs.
- **Apache Solr** is the popular, blazing-fast, open source enterprise search platform built on Apache Lucene.
- **Apache Sqoop** is a tool designed for efficiently transferring bulk data between Apache Hadoop and structured data stores such as relational databases.
- **Apache Storm** is a free and open source distributed real-time computation system.

Big Data Platform:

- **Snowflake Elastic Data Warehouse** A data warehouse that is more flexible, scalable, and easy to use than anything else available.
- **Domo** brings the business and all its data together in one intuitive platform.
- **Attivio 5.0** unlocks the business value trapped in text-based sources of information by making it easy to analyze dark data – giving you a more complete view so you can act with certainty.
- **Data Torrent RTS 3**, which runs on Amazon EMR, is powering PubMatic's real-time Ad analytics platform enabling publishers to drive the highest value for their digital media assets.

- **MongoDB 3.0** features performance and scalability enhancements that place MongoDB at the forefront of the database market as the standard DBMS for modern applications.
- **CouchBase N1Ql** is the first query language to leverage the complete flexibility of JSON with the full power of SQL.
- **HP Vertica Excavator** is the latest version of HP Vertica enables organizations to quickly ingest and analyze high-speed streaming data, from various sources, including Internet of Things applications, and provides enhanced SQL analytics and performance to Hadoop.

Big Data Visulatization Tool:

- **Datameer** is an end-to-end big data discovery and analytics platform, purpose built for Hadoop.
- **GoodData** is a cloud SaaS analytics platform which connects to a large number of data sources – big data, databases, online apps, social data, and so on.
- **Lumify** is an open source project to create a big data fusion, analysis, and visualization platform designed for anyone to use.
- **Spotfire** supports easy to build data visualization, text analytics, predictive analytics and statistical analysis.

Big Data Analytics:

The primary goal of big data analytics is to help companies make more informed business decisions by enabling data scientists, predictive modelers and other analytics professionals to analyze large volumes of transaction data, as well as other forms of data that may be untapped by conventional business intelligence (BI) programs. That could include Web server logs and Internetclickstream data, social media content and social network activity reports, text from customer emails and survey responses, mobile-phone call detail records and machine data captured by sensors connected to the Internet of Things. Big data can be analyzed with the software tools commonly used as part of analytics disciplines such as predictive analytics, data mining, text analytics and statistical analysis. Mainstream BI software and data visualization tools can also play a role in the analysis process.

Many organizations looking to collect, process and analyze big data have turned to a newer class of technologies that includes Hadoop and related tools such as YARN, MapReduce, Spark, Hive and Pig as well as NoSQL databases. Those technologies form the core of an open source software framework that supports the processing of large and diverse data sets across clustered

Hadoop clusters and NoSQL systems are being used as landing pads and staging areas for data before it gets loaded into a data warehouse for analysis, often in a summarized form that is more conducive to relational structures. Increasingly though, big data vendors are pushing the concept of a Hadoop data lake that serves as the central repository for an organization's incoming streams of raw data.

Major Trends In Big Data Analytics:

1. Big data analytics in the cloud :

Hadoop, a framework and set of tools for processing very large data sets, was originally designed to work on clusters of physical machines. That has changed. "Now an increasing number of technologies are available for processing data in the cloud," says Brian Hopkins, an analyst at Forrester Research. Examples include Amazon's Redshift hosted BI data warehouse, Google's BigQuery data analytics service, IBM's Bluemix cloud platform and Amazon's Kinesis data processing service. "The future state of big data will be a hybrid of on-premises and cloud.

2. Hadoop: The new enterprise data operating system :

Distributed analytic frameworks, such as MapReduce, are evolving into distributed resource managers that are gradually turning Hadoop into a general-purpose data operating system, with these systems; we can perform many different data manipulations and analytics operations by plugging them into Hadoop as the distributed file storage system.

Hadoop with adequate performance, more businesses will use Hadoop as an enterprise data hub. "The ability to run many different kinds of [queries and data operations] against data in Hadoop will make it a low-cost, general-purpose place to put data that you want to be able to analyze.

3. Big data lakes :

A data lake, also called an enterprise data lake or enterprise data hub and it take the data sources and dump them all into a big Hadoop repository .It provides tools for people to analyze the data, along with a high-level definition of what data exists in the lake. People build the views into the data as they go along. It's a very incremental, organic model for building a large-scale database.

4. More predictive analytics:

Traditional machine learning uses statistical analysis based on a sample of a total data set. Now we have the ability to do very large numbers of records and very large numbers of attributes per record" and that increases predictability. To enable real-time analysis and predictive modeling out of the same Hadoop core, a large-scale data processing engine, and it's associated SQL query tool, Spark SQL. It has fast interactive query as well as graph services and streaming capabilities.

5. SQL on Hadoop: Faster, better:

These tools are nothing new. Apache Hive has offered a structured, SQL-like query language for Hadoop for some time. But commercial alternatives from Cloudera, Pivotal

Software, IBM and other vendors not only offer much higher performance, but also are getting faster all the time. That makes the technology a good fit for "iterative analytics," where an analyst asks one question, receives an answer, and then asks another one. That type of work has traditionally required building a data warehouse. SQL on Hadoop isn't going to replace data warehouses, at least not anytime soon, but it does offer alternatives to more costly software and appliances for certain types of analytics

6. More, better NoSQL:

Alternatives to traditional SQL-based relational databases, called NoSQL databases, are rapidly gaining popularity as tools for use in specific kinds of analytic applications. There are 15 to 20 open-source NoSQL databases out there, each with its own specialization. For example, a NoSQL product with graph database capability, such as ArangoDB, offers a faster, more direct way to analyze the network of relationships between customers or salespeople than does a relational database.

7. Deep learning:

Deep learning, a set of machine-learning techniques based on neural networking, is still evolving but shows great potential for solving business problems; Deep learning . . . enables computers to recognize items of interest in large quantities of unstructured and binary data, and to deduce relationships without needing specific models or programming instructions

8. In-memory analytics:

The use of in-memory databases to speed up analytic processing is increasingly popular and highly beneficial. In fact, many businesses are already leveraging hybrid transaction/analytical processing (HTAP) — allowing transactions and analytic processing to reside in the same in-memory database. We can perform analytics faster with HTAP, all of the transactions must reside within the same database.Big Data has seen a huge leap forward in 2015 regarding how it has been represented and used across companies.

In-Memory Databases:

As the use of data has increased in the past year, the speed at which results are needed has grown with it. When this hasn't been the case, people want to be more informed than before or have the ability to make decisions in real time, rather than through the use of reports reporting on historical data.

Memory databases allow companies the freedom to access, analyze and take actions based on data much quicker than regular databases. This in turn means that either

decision can be made quicker as data can be analyzed faster or more informed as more data can be analyzed in the same amount of time.

Non-Data Scientists:

Big data having automated platforms that can allow employees who may not have as much skill with data as others, to collect, analyze and make decisions based on this data. It main tasks helps to create business results.

More Sensor Driven Data:

It contains sensor-to-sensor data being collected, collated and analyzed through purely sensor based collection. This can be done in multiple ways from the way that objects are interacting with another object, to the settings that people are using on particular devices.

Sensor based data is unlikely to out perform transactional data (which currently makes up the majority of data collection) but is still likely to see a marked increase. Which could see this number grow beyond transactional data within the next 5 years?

Deeper Customer Insight:

It designed in multi-dimensional ways to create even deeper customer insight.

.With this new technology allows metrics to be tracked across even more areas and wearable creating even more possible trackable actions, deeper customer understanding is inevitable.

HR Analytics

In 2015 more companies had come up with an effective HR strategy and optimizing workflow to track overall employee happiness, we are likely to see an increased use of HR analytics in 2015 as those who have adopted it within the last couple of years will be gaining ground on their competitors and those who don't have it will need to catch up.

The adoption rates have grown and the importance of Big Data as a business function has increased, but what are we going to see in 2016?

1. Big data gets cloudy, ETL gets personal and NOSQL is on the rise :

The cloud is everywhere, and we will continue to see adoption at extreme volumes. And big data is driving a lot of cloud growth: It has been suggested that 80 percent of an analyst's time is spent on data prep, while only 20 percent is spent looking for insights. Tools like Trifocal, Alteryx, Paxata and Informatica Rev are making data preparation easier to use with less technology and infrastructure required to support it.

2. Big data will present significant challenges with regard to securing sensitive information and protecting individual privacy, and the only answer will be open technologies, standardized interconnectivity and collaboration on a global scale.
3. The adoption of the enterprise data hub concept will result in immediate enhancements to developer agility. Result: There will soon be two types of companies: those using big data to dominate in the market –and those being dominated by others
4. By mid-2016 we expect to see using enterprise data hub approaches to enhance agility and decision-making that leads to market dominating strategies.

Conclusion

Big Data is emerging from the realms of science projects at Web companies to help companies like telecommunication giants understand exactly which customers are unhappy with service and what processes caused the dissatisfaction, and predict which customers are going to change carriers. To obtain this information, billions of loosely-structured bytes of data in different locations need to be processed until the needle in the haystack is found. The analysis enables executive management to fix faulty processes or people and maybe be able to reach out to retain the at-risk customers. The real business impact is that big data technologies can do this in weeks or months, four-or-more-times faster than traditional data warehousing approaches. This paper also focused on future and emerging trends which create particular opportunities, but also challenges for the software and services industry. Big Data actors can create a better momentum in terms of global competition. However, many technical challenges described in this paper must be addressed before this potential can be realized fully. These challenges will require transformative solutions, and will not be addressed naturally by the next generation of industrial products. We must support and encourage fundamental research towards addressing these technical challenges if we are to achieve the promised benefits of Big Data.

References

1. The Great Cloud Migration: Your Roadmap to Cloud Computing, Big Data and Linked Data by Michael C.Daconta
2. Big data: The next frontier for innovation, competition, and productivity. James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. McKinsey Global Institute.
3. Planning for Big Data by O'Really Radar Team by Edd Dumbill
4. Big Data: Using SMART Big Data, Analytics and Metrics to Make Better Decisions and Improve Performance by Bernard Marr
5. Big Data: A Revolution That Will Transform How We Live, Work and Think by Viktor Mayer schonberger and Kenneth Cukier