BIG DATA – A SURVEY OF BIG DATA TECHNOLOGIES

Pandya Dhavalchandra,
Atmiya Institute of Technology and Science (AITS), Rajkot, Gujarat, India

Mahidhareeya Jignasu,
Atmiya Institute of Technology and Science (AITS), Rajkot, Gujarat, India

Raval Amit
Atmiya Institute of Technology and Science (AITS), Rajkot, Gujarat, India

Abstract— In this paper, we reviewed the background of big data current trends and technologies. We first introduce the general definition of big data and review related technologies and techniques, such as cloud computing, Internet of Things, data centers, and Hadoop. Many business cases exploiting big data have been realized in recent years; Twitter, LinkedIn, and Facebook are examples of companies in the social networking domain. Other big data use cases have focused on capturing of value from streaming of movies (Netflix), monitoring of network traffic, or improvement of processes in the manufacturing industry. After that we discuss four V's of Big Data, for each phase, we introduce the background, discuss the current techniques and technologies. We finally examine the several representative applications of big data, including Internet of Things, online social networks, medial applications, collective intelligence, and smart grid.

Keywords— Big Data, Big Data Definition, Association Analysis, Data Centre, Big Data Analysis

I. INTRODUCTION

Big Data is the ocean of information we swim in every day. Vast zeta bytes of data flowing from our computers, mobile devices and machine sensors. With the right solutions, organizations can dive into all that data and gain valuable insights that were previously unimaginable. [22]

Big Data is Broad Term for data sets so large or complex that traditional data processing applications are inadequate. Challenges include analysis, capture, data creation, search, sharing, storage, transfer, visualization, and querying and information privacy. [23]

Over the past 20 years, data has increased in a large scale in various fields. According to a report from International Data Corporation (IDC), in 2011, the overall created and copied data volume in the world was 1.8ZB (10^21 B), which increased by nearly nine times within five years [1]. This figure will double at least every other two years in the near future [1].

Nowadays, big data related to the service of Internet companies grow rapidly. For example, Google processes data of hundreds of Petabyte (PB), Facebook generates log data of over 10 PB per month, Baidu, a Chinese company, processes data of tens of PB, and Taobao, a subsidiary of Alibaba generates data of tens of Terabyte (TB) for online trading per day [2].

II. SURVEYED TECHNOLOGIES

Major surveyed technologies have been summarized as shown in Fig.-1. It is desirable to consider all possible areas to check the current trends and future possibilities. This paper collects information from major possible areas to make the overall survey.

III. TOOLS AND TECHNOLOGIES

This section shows major trends and tools from the review of studied literatures.

A. Big Data Analysis Platforms and Tools

1) Hadoop: Hadoop is an open source framework for storing and processing large datasets using clusters of commodity hardware. Hadoop is designed to scale up to hundreds and even thousands of nodes and is also highly fault tolerant [3]. An open source (free) software framework for processing huge datasets on certain kinds of problems on a distributed system. Its development was inspired by Google's MapReduce and Google File System. It was originally developed at Yahoo! and is now managed as a project of the Apache Software Foundation [4].

2) Hadoop distributed file system (HDFS): HDFS is applied to store smart meter data and weather data [5]. HDFS is a distributed file system that is used to store data across cluster of commodity Machines while providing high availability and fault tolerance [6].

3) Hadoop Map Reduce: MapReduce is a software framework introduced by Google for processing huge datasets on certain kinds of problems on a distributed system. Also implemented in Hadoop [4]. MapReduce in Hadoop processes parallelizable problems across huge datasets by using a large number of computers (nodes). The process includes three steps: map, shuffle, and reduce [5]. MapReduce is a simple but powerful programming model for large-scale computing using a large number of clusters of commercial PCs to achieve automatic parallel processing and distribution. In MapReduce, computing model only has two functions, i.e., Map and Reduce, both of which are programmed by users [2]. The MapReduce framework deals with data mapped on distributed file systems, with intermediate data being stored on local disks and can be retrieved remotely by reducers. Google's proprietary MapReduce paradigm reads and writes to the Google File System [7]. The programming model used in Hadoop is MapReduce which was proposed by Dean and Ghemawat at Google. MapReduce is the basic data processing scheme used in Hadoop which includes breaking the entire task into two parts, known as mappers and reducers [8].

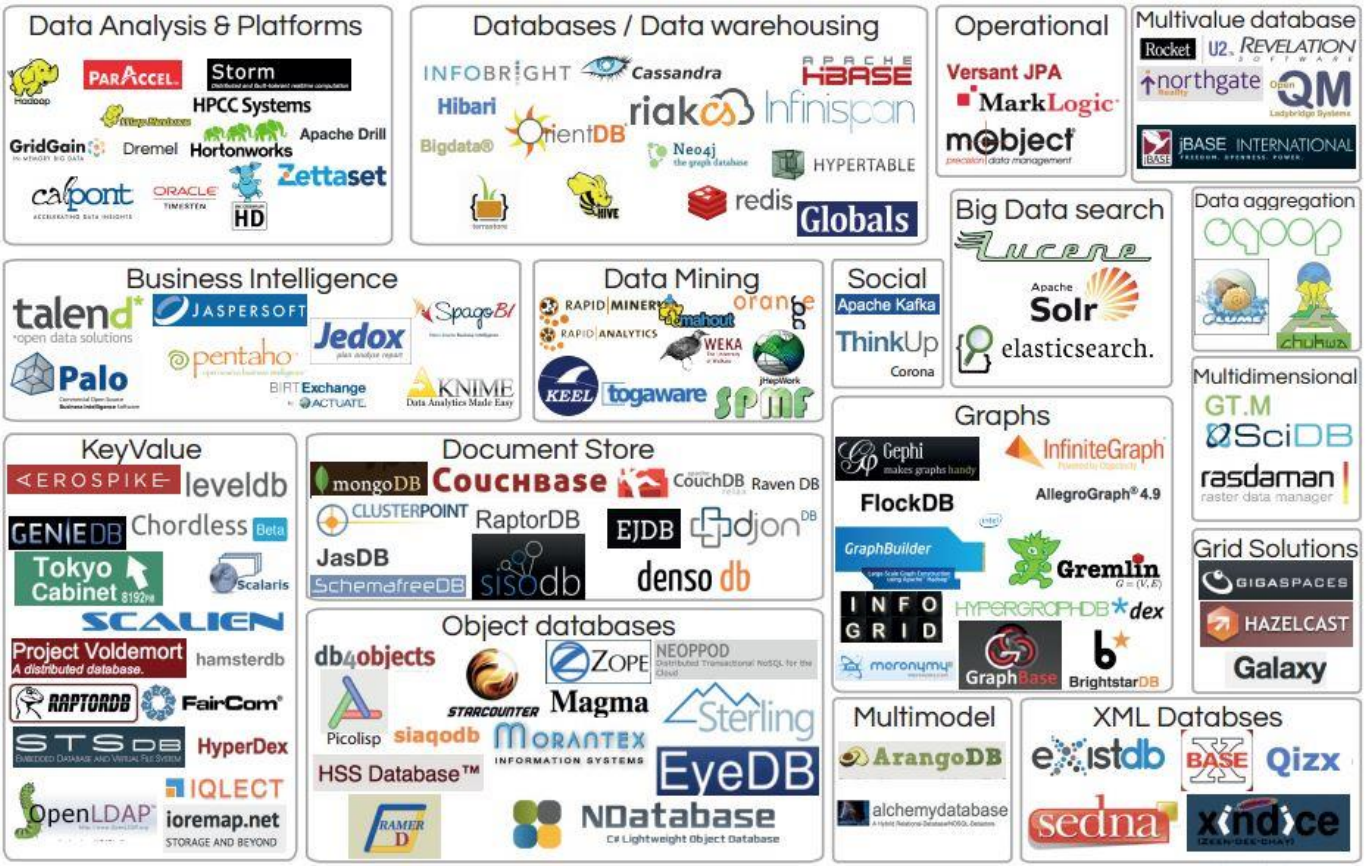

Fig. 1 [23]

4) NoSQL: Traditional relational databases cannot meet the challenges on categories and scales brought about by big data. NoSQL databases (i.e., non-traditional relational databases) are becoming more popular for big data storage. NoSQL databases feature flexible modes, support for simple and easy copy, simple API, eventual consistency, and support of large volume data. NoSQL databases are becoming the core technology for of big data [2]. Need for databases being distributed and horizontally scalable. "NoSQL Definition: Next Generation Databases mostly addressing some of the points: being non-relational, distributed, open-source and horizontally scalable. The original intention has been modern web-scale databases. The movement began early 2009 and is growing rapidly. Often more characteristics apply such as: schema-free, easy replication support, simple API, eventually consistent / BASE (not ACID), a huge amount of data and more [9]. NoSQL databases have been increasingly used to overcome the inflexibility of relational databases with regard to highly heterogeneous data, and to provide improved support for distributed queries and integrated caching (Xiang and Hou, 2010). NoSQL databases do not have a predefined schema that dictates a uniform and fixed definition of the stored data in rows. In this way, database fields can be modified over time and can adapt to future requirements [10].

5) Data Warehouses: Specialized database optimized for reporting, often used for storing large amounts of structured data. Data is uploaded using ETL (extract, transform, and load) tools from operational data stores, and reports are often generated using business intelligence tools [4].

6) Amazon S3: Amazon S3 (Simple Storage Service) is an online file storage web service offered by Amazon Web Services. Amazon S3 provides storage through web services interfaces (REST, SOAP, and BitTorrent). [23]

7) Apache Flume: Flume is a distributed, reliable,

and available service for efficiently collecting, aggregating, and moving large amounts of log data. It has a simple and flexible architecture based on streaming data flows. It is robust and fault tolerant with tunable reliability mechanisms and many failover and recovery mechanisms. [24]

8) Data Mart: Subset of a data warehouse, used to provide data to users usually through business intelligence tools [4].

9) Terracotta: BigMemory from Terracotta enables a distributed in-memory data management solution [11].

10) Apache Oozie: Oozie is a workflow scheduler system to manage Apache Hadoop jobs. Oozie is integrated with the rest of the Hadoop stack supporting several types of Hadoop jobs out of the box (such as Java map-reduce, Streaming map-reduce, Pig, Hive, Sqoop and Distcp) as well as system specific jobs (such as Java programs and shell scripts).[25]

11) Yarn: Hadoop YARN is a resource management layer and schedules the jobs across the cluster [12].

12) Jaspersoft: Jaspersoft BI Enterprise is an analytics platform for enterprise domain, and consists of several functionalities (ETL, OLAP, in-memory server, visualization) [11]

13) Pentaho: Pentaho is one of the most popular open-source BI software. It includes a web server platform and several tools to support reporting, analysis, charting, data integration, and data mining, etc., all aspects of BI. Weka’s data processing algorithms are also integrated in Pentaho and can be directly called [2].

14) Talend Open Studio: Talend Open Studio for Data Integration is an open source data integration product developed by Talend and designed to combine, convert and update data in various locations across a business. [23]

15) Karmasphere: Web browser based UI capabilities, Physical HW and cloud Execution environment, Delimited (CSV, TSV, XML etc.), text, extended text, sequence, binary Data sources, Analytical functions, batch analysis (Hadoop) Data analysis capability [11]

IV. SRVERY TABLE

Sr. No.	Big Data Analysis Platforms and Tools	Sr. No.	Big Data Databases	Sr. No.	Big Data Mining Tools	Sr. No.	Big Data File Systems and Programming Languages	Sr. No.	Big Data Tools - Transfer and Aggregate
1	Hadoop	1	Cassandra	1	RapidMiner	1	Gluster	1	Lucene
2	Hadoop distributed file system (HDFS)	2	HBase	2	Mahout	2	Hadoop Distributed File System	2	Solr
3	Hadoop Map Reduce	3	MongoDB	3	Orange	3	Pig	3	Sqoop
4	NoSQL	4	SimpleDB	4	Weka	4	R	4	Flume
5	Data Warehouses	5	Neo4j	5	DataMelt	5	KNIME	5	Chukwa
6	Amazon S3	6	CouchDB	6	KEEL	6	ECL		
7	Apache Flume	7	OrientDB	7	SPMF				

8	Data Mart	8	Terrastore	8	Rattle				
9	GridGain	9	FlockDB						
10	HPCC Systems	10	Hibari						
11	Storm	11	Riak						
12	Terracotta	12	Hypertable						
13	Avro	13	Blazegraph						
14	Apache Oozie	14	Hive						
15	Yarn	15	InfoBright Community Edition						
16	Jaspersoft	16	Infinispan						
17	Pentaho	17	Redis						
18	Talend Open Studio								
19	Zookeeper								
20	Wibidata								
21	Karmasphere								
22	SkyTree								

Table-1

B. Big Data Databases

1) Cassandra: Cassandra is a distributed storage system to manage the huge amount of structured data distributed among multiple commercial servers [13]. The system was developed by Facebook and became an open source tool in 2008. It adopts the ideas and concepts of both Amazon Dynamo and Google BigTable, especially integrating the distributed system technology of Dynamo with the BigTable data model [2].

2) HBase: An open source (free), distributed, non-relational database modeled on Google's Big Table. It was originally developed by Powerset and is now managed as a project of the Apache Software foundation as part of the Hadoop [4].

3) MongoDB: MongoDB is open-source and document-oriented database [14]. MongoDB stores documents as Binary JSON (BSON) objects [15], which is similar to object. Every document has an ID field as the primary key. Query in MongoDB is expressed with syntax similar to JSON [2].

4) SimpleDB: SimpleDB is a distributed database and is a web service of Amazon [16]. Data in SimpleDB is organized into various domains in which data may be stored, acquired, and queried. Domains include different properties and name/value pair sets of projects. Date is copied to different machines at different data centers in order to ensure data safety and improve performance [2].

5) CouchDB: Apache CouchDB is a documentoriented database written in Erlang [17]. Data in CouchDB is organized into documents consisting of fields named by keys/names and values, which are stored and accessed as JSON objects. Every document is provided with a unique identifier. CouchDB allows access to database documents through the RESTful HTTP API [2].

6) OrientDB: OrientDB had best perfor-mance, when compared to the alternatives (AllegroGraph, Fuseki, Neo4j, Titan) [18].

7) Hypertable: HyperTable was developed similar to BigTable to obtain a set of high-performance, expandable, distributed storage and processing systems for structured and unstructured data [19]. HyperTable relies on distributed file systems, e.g. HDFS and distributed lock manager. Data representation, processing, and partition mechanism are similar to that in BigTable.[2]

8) Hive: Hive for data summarization and ad hoc querying [20]. Hive used for MapReduce [2].

Hive is another MapReduce wrapper developed by Facebook. This wrapper provide a better environment and make the code development simpler since the programmers do not have to deal with the complexities of MapReduce coding [3].

C. Big Data Mining Tools

1) RapidMiner: Rapidminer is an open source software used for data mining, machine learning, and predictive analysis. In an investigation of KDnuggets in 2011, it was more frequently used than R (ranked Top 1). Data mining and machine learning programs provided by RapidMiner include Extract, Transform and Load (ETL), data pre-processing and visualization, modeling, evaluation, and deployment [2].

D. Big Data File Systems and Programming Languages

1) Pig: Apache Pig is a SQL-like environment developed at Yahoo is being used by many organizations like Yahoo, Twitter, AOL, LinkedIn etc. This wrapper provide a better environment and make the code development simpler since the programmers do not have to deal with the complexities of MapReduce coding [3].

2) R: R, an open source programming language and software environment, is designed for data mining/analysis and visualization. While computing intensive tasks are executed, code programmed with C, C++ and Fortran may be called in the R environment. In addition, skilled users can directly call R objects in C. Actually, R is a realization of the S language, which is an interpreted language developed by AT&T Bell Labs and used for data exploration, statistical analysis, and drawing plots.[2]

3) KNIME: KNIME (Konstanz Information Miner) is a user-friendly, intelligent, and open-source rich data integration, data processing, data analysis, and data mining platform [21]

V. **CONCLUSIONS**

This paper reported on the current technologies and tools of Big Data research by examining the literature and identifying current trends. Big Data improves the lives using Big Storage, Analytics and Clouds, which are applicable to save the time and money with more feasibility.

Major Industrial and Environmental areas are using Big Data technologies now. Survey Table-1 shows the available technologies and tools.

VI. **REFERENCES**

[1] Gantz J, Reinsel D (2011), Extracting value from chaos, IDC iView, pp 1–12

[2] M. Chen, S. Mao, Y. Liu, Big Data: A Survey, DOI 10.1007/s11036-013-0489-0

[3] Dilpreet Singh and Chandan K. Reddy, A Survey on Platforms for Big Data Analytics, Journal of Big Data, 1:1, 8, 2014

[4] Ebook - Big data: The next frontier for innovation, competition, and productivity

[5] Pei Zhang, Member, CSEE, Senior Member, IEEE, Xiaoyu Wu, Student Member, IEEE, Xiaojun Wang, and Sheng Bi, Short-Term Load Forecasting Based on Big Data Technologies, CSEE JOURNAL OF POWER AND ENERGY SYSTEMS, VOL. 1, NO. 3, SEPTEMBER 2015

[6] Borthakur D (2008), HDFS architecture guide, HADOOP APACHE PROJECT. http://hadoop.apache.org/docs/r1.2.1/hdfs_design.pdf

[7] Jyotsna Talreja Wassan, Discovering Big Data Modelling for Educational World, Procedia - Social and Behavioral Sciences 176 (2015) 642 – 649

[8] Lee K-H, Lee Y-J, Choi H, Chung YD, Moon B (2012), Parallel data processing with MapReduce: a survey, ACM SIGMOD Record 40(4):11–20

[9] Prof. Dr. Uta Stor, Big Data Technologies

[10] Claudia Vitolo, Yehia Elkhatib, Dominik Reusser, Christopher J.A. Macleod, Wouter Buytaert, Web technologies for environmental Big Data, Environmental Modelling & Software 63 (2015) 185e198

[11] PekkaPaakkonen, DanielPakkala, Reference Architecture and Classification of Technologies, Products and Services for Big Data Systems, Big Data Research 2 (2015) 166–186

[12] Vavilapalli VK, Murthy AC, Douglas C, Agarwal S, Konar M, Evans R, Graves T, Lowe J, Shah H, Seth S (2013), Apache hadoop yarn: Yet another resource negotiator, In: Proceedings of the 4th annual Symposium on Cloud Computing., p 5

[13] Lakshman A, Malik P (2009), Cassandra: structured storage system on a p2p network, In: Proceedings of the 28th ACM symposium on principles of distributed computing. ACM, pp 5–5

[14] Chodorow K (2013), MongoDB: the definitive guide, O'Reilly Media Inc

[15] Crockford D (2006), The application/json media type for javascript object notation (json)

[16] Murty J (2009) Programming amazon web services: S3, EC2, SQS, FPS, and SimpleDB. O'Reilly Media Inc

[17] Anderson JC, Lehnardt J, Slater N (2010) CouchDB: the definitive guide. O'Reilly Media Inc

[18] M. Dayarathna, T. Suzumura, Graph database benchmarking on cloud environ-ments with XGDBench, Autom. Softw. Eng. 21 (2013).

[19] Judd D (2008) hypertable-0.9. 0.4-alpha

Hsinchun Chen, Roger H. L. Chiang, Veda C. Storey, BUSINESS INTELLIGENCE AND ANALYTICS: FROM BIG DATA TO BIG IMPACT, MIS Quarterly Vol. 36 No. 4, pp. 1165-1188/December 2012

[20] Berthold MR, Cebron N, Dill F, Gabriel TR, K¨otter T, Meinl T, Ohl P, Sieb C, Thiel K, Wiswedel B (2008), KNIME: the Konstanz information miner, Springer

[21] http://go.sap.com/solution/big-data.html

[22] www.wikipedia.com

[23] https://flume.apache.org/

[24] https://oozie.apache.org/

[25] Image:https://sranka.wordpress.com/2014/01/29/ big-data-technologies/