ORIENTAL JOURNAL OF
COMPUTER SCIENCE & TECHNOLOGY
An International Open Free Access, Peer Reviewed Research Journal
Published By: **Oriental Scientific Publishing Co., India.**
www.computerscijournal.org

ISSN: 0974-6471
April 2016,
Vol. 9, No. (1):
Pgs. 52-56

Storage and Computation on Big Data: A Comparative Study

CH SRAVAN KUMAR[1], P. BUDDHA REDDY[2] and K. SRINIVAS[3]

Assistant Professor, Vardhaman College of Engineering, Hyderabad, India.
*Corresponding author E-mail: chakralasravan@gmail.com

http://dx.doi.org/10.13005/ojcst/901.10

(Received: February 16, 2016; Accepted: March 02, 2016)

ABSTRACT

A huge data space includes set of interesting points; Skyline is an important operation in many applications to return a set of interesting points from a potentially huge data space [1]. This survey paper highlights the characteristics of big data and their challenges. This paper also discusses the tools and techniques of big data. The existing algorithms like SaLSa, SSPL are novel computation algorithms. SaLSa exploits the idea of presorting the input data so as to effectively limit the number of tuples to be read and compared [2]. SSPL utilizes sorted positional index lists which require low space overhead to reduce I/O cost significantly [1]. SSPL consists of two phases. In phase 1, SSPL computes scan depth of the involved sorted positional index lists. During retrieving the lists in a round-robin fashion, SSPL performs pruning on any candidate positional index to discard the candidate whose corresponding tuple is not skyline result. Phase 1 ends when there is a candidate positional index seen in all of the involved lists. In phase 2, SSPL exploits the obtained candidate positional indexes to get skyline results by a selective and sequential scan on the table [1].

Keywords: SaLSa, SSPL, Big Data.

INTRODUCTION

As data been increased in everyday life, computing the data and keeping the data secure is one of the important parameters. Over the years, disk storage capacity has improved from 40 MB to 3TB, while disk transfer rate has also been improved from 800 KB/s to 600 MB/s. Here we are indexing two computation algorithms (SaLSa, SSPL) on data sets.

Formats of Big Data and its resources

A huge collection of data is called as Big data. The collection of data over a time frame that is so complex and difficult to process and manage using conventional database management tools[2]. Big Data and its sources can be categorized into following categories:

Semi-structured Data

Such as XML formatted data.

Unstructured Data

These data can be generated by humans such as social media, discussion forums and customer feedback, comments, emails etc. or may be generated by machine such as online transactional, satellite and environmental data collected through various sensors, web-logs, call records etc.

Structured Data

Generated from various researches efforts, CRM (Customer Relationship Management) and other such traditional databases.

Big Data Characteristics and Big Data Challenges

There are three basic characteristics of Big Data are: Volume, Variety and Velocity. Each characteristic has a challenge in handling and processing the data. These challenges could be in storage, sorting, searching, collection, integration, analysis, retrieval, and visualization from the various mentioned key concepts of the Big Data.

Variety

The data is being collected from multiple sources in different formats already discussed - Structured data, semi-structured data and unstructured data. Out of which the unstructured data is a big hurdle in computing and analysis part as they do not have a common format, therefore a common tool or algorithm cannot be followed in variety of modalities of the data.

Volume

As per current scenario, various sources of data generations throughout the world, generating the data at tremendous speed per day. Facebook and Twitter are the kind of social media that produce daily approximately 500 TB and 7 TB of data respectively. According to a survey done by IBM, 2.5 quintillion bytes of data are being generated every day. A quintillion equals 1018 bytes.

Velocity

This aspect of Big Data is associated with the speed at which data is being produced and processed. When we look for the real time processing and response the speed of data production becomes a critical challenge for analytical and visualization tools. If the response time of the analytical tools is not capable to cope up with speed of data arriving, the result becomes useless.

Handling and Processing of Big Data [3]

The following five steps are used to handle and process the big data:

Take a bird's eye view

You need a good overview of what big data your organization has – so conduct a consumer-centric data audit. IT should definitely play a part as the department is likely to be most at home with data, while data analysts should also be involved as they are familiar with combining and using disparate data sources. The privacy or legal team should also be engaged to ensure regulations are adhered to. Produce a comprehensive list of the raw materials you have available for any big data initiatives and identification of the gaps where the data is unavailable.

The customer is king

Big data creates new opportunities. Every marketer knows that for a brand to be successful, it has to have a compelling offer delivered to consumers via the right channel at the right time. Ensure you're aligning your efforts to your business objectives. For marketing and insight teams this typically means focusing on initiatives that benefit your consumers.

Roll out a roadmap

You're now ready to build on the results of the test. Your roadmap should outline how any proof of concept tests can be operationalised and define future tests. You should always be on the lookout for new, significant data. Priority consideration should be given to what big data can most quickly be captured and translated into these existing environments.

Get on your marks

You must now build a strategy for managing and making big data actionable. You should analyze raw data to determine when and how it can be used; make data operational quickly and efficiently; identify, and if necessary discard, junk data which will clog the system; automate decisioning to accommodate the variety and velocity

of Big Data; and carry out all work in a way that is compliant with current data laws.

Set off on a test run

A significant investment may be required to establish a big data environment, not just in terms of hardware but also human resource. Demonstrating ROI is therefore crucial to securing sponsorship from the business. It should be perfectly possible to execute big data use cases without building a new, full-blown environment for doing so. Activities at this stage include statistical analysis (mining), searching for predictive patterns and attempting to turn these into processes which can be tested in real-world scenarios.

Organization of Paper

The organization of the paper is as follows: First, introduces the Big Data, formats of big data and resources, their characteristics. Also, it discusses the need of handling and processing of Big Data in current scenario in different areas of applications. Second, a description about four well known tools and techniques for storing and four for computing Big Data with along with their advantages/ disadvantages and the suitable environment they are applicable to work with. Third, gives the comparison of various tools and techniques based on their capabilities and limitations associated with them. Fourth, gives finally concludes this paper with some useful suggestions and recommendations.

Big Data Framework

Four Big Data strategies [4]

- Performance Management
- Data Exploration
- Social Analytics
- Decision Science

With respect to future trends in the Big Data field, the following practices are starting to emerge:

- Integrating multiple big data strategies.
- Build a Big Data capability.
- Be proactive and create a Big Data policy.

Comparison of various tools and techniques

Data Wrangler

Wrangler is an interactive tool for data cleaning and transformation[4] Wrangler is designed to quicken the process of data manipulation, helps u spend less time transforming your data and more time learning from it. Wrangler allows interactive transformation of messy, realworld data into the data tables which analysis tools expect which in turns allows spending less time formatting and more time analyzing your data.

The R Project

R is a free software environment for statistical computing and graphics. R is a general statistical analysis platform that runs on the

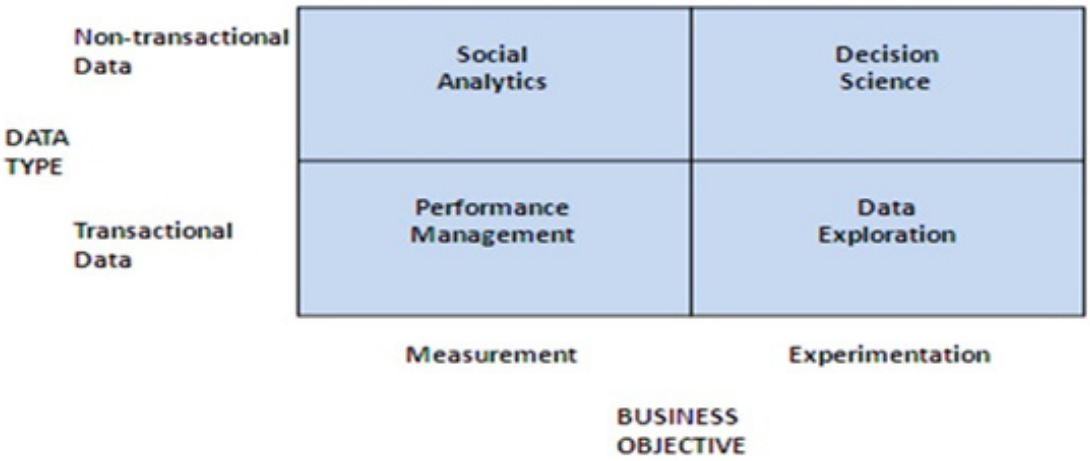

Fig. 1: Big Data Framework[4]

command line[4] R can find means, medians, standard deviations, correlations and much more, including linear and generalized linear models, nonlinear regression models, time series analysis, classical parametric and nonparametric tests, clustering and smoothing.

Time Flow

This is desktop software for analyzing the time attribute. TimeFlow can generate visual timelines from text files, with entries color- and size-coded for easy pattern spotting. It also allows the information to be sorted and filtered, and it gives some statistical summaries of the data.

NodeXL

NodeXL is visualization and analysis software of networks and relationships. It uses a technology referring to the discipline of finding connections between people based on various data sets

Table 1: Popular Big Data Techniques [4]

Transactional Data	Technique Business Intelligence (BI)/Online Analytical Processing (OLAP): •Users interactively analyze multidimensional data •Users can roll-up, drill-down, and slice data •BI tools provide dashboard and report capabilities Cluster Analysis: •Segment objects (e.g., users) into groups based on similar properties or attributes Data Mining: •Process to discover and extract new patterns in large data sets Predictive Modeling: •A model is created to best predict the probability of an outcome SQL: •A computer language that manages (e.g., query, insert, delete, extract) A/B Testing: •A method of testing in which a control group is compared to test groups to determine if there is an improvement based on the test condition •Often used in website design to test for higher conversion rates
Non-transactional Data	Crowdsourcing: •A process for collecting data from a large community or distributed group of people •Idea submission is a common crowdsourcing activity Textual Analysis: •Computer algorithms that analyze natural language •Topics can be extracted from text along with their linkages Sentiment Analysis: •A form of textual analysis that determines a positive, negative, or neutral reaction •Often used in marketing brand campaigns Network analysis: •A methodology to analyze the relationship among nodes (e.g., people) •On social media platforms, it can be used to create the social graph of follower and friends' connections among users

Table 2: Comparison of Various Tools for Data Storage

Tool	Category	Multipurpose visualization	Mapping	Data Stored or Processed?	Designed for Web Publishing
DataWrangler	Data Cleaning	No	No	External l Server	No
R Project	Statistical Analysis	Yes	With plug-in	Local	No
TimeFlow	Temporal data analysis	No	No	Local	No
NodeXL	Network analysis	No	No	Local	As image
CSVKit	CSV file analysis	No	No	Local or external server	Yes
Tableau	Visualization app/service	Yes	Yes	Public external server	Yes

Tableau Plateau

Data visualization tools allow anyone to organize and present information intuitively. It is exceptionally powerful in business because it communicates insights through data visualization[8] This tool can turn data into any number of visualizations, from simple to complex. You can drag and drop fields onto the work area and ask the software to suggest a visualization type, then customize everything from labels and tool tips to size, interactive filters and legend display [7].

CSV Kit

CSVKit contains tools for importing, analyzing and reformatting comma-separated data files. CSVKit makes it quick and easy to preview, slice and summarize your file to examine it[6].

CONCLUSION

This Paper presents the various characteristics of big data and their tools and techniques and their comparison studies and with a framework to all the components of Big Data.

REFERENCES

1. Xixian, Han., Jianzhong, Li., "Efficient Skyline Computation on Big Data," IEEE transactions on knowledge and data engineering, **25**(11): 2013.
2. I., Bartolini, P., Ciaccia, M., Patella, "Efficient Sort-Based Skyline Evaluation," *ACM Trans. Database Systems,* **33**(4): pp. 31:1-31:49, 2008.
3. bigdataweek.com/blog/2013/04/08/five-steps-to-handling-big-data/
4. http://iveybusinessjournal.com/publication/four-strategies-to-capture-and-create-value-from-big-data/
5. Sofiya Mujawar., Aishwarya Joshi., "Data Analytics Types, Tools and their Comparison", *ijarcce,* **4**(2): (2015).
6. http://www.dataversity.net/3-types-data-analytics-descriptivepredictive-prescriptive
7. http://www.computerworld.com
8. http://www.kdnuggets.com/2014/06/top-10-data-analysis-toolsbusiness.html