Designing database to efficiently store and retrieve massive amount of data

Guided Research Methods

Nurlan Imanov 31.07.2023

Project objective

- **The main objective to answer this question:** *The company has started to store its textual data on RDBMS. The project gets big scale and they face database performance issues. The database migration(from RDBMS to NoSQL) can be costly. They seek the solid reason and investigation of switching from RDBMS to NoSQL. They also want to see the limitation of current RDBMS. Thus, the optimization of RDBMS is needed first to see the performance metrics. In order to see the performance of NoSQL, tests has to be conducted on one of the NoSQL database in distributed environment. As a conclusion the company has to find the answer of this question: which database design is needed to meet the requirements of future needs in terms of Big Data . Therefore reasonable metrics has to be compared. The company predicts to have Big Data issues in near future. For that reason this experiments has to be analyzed in terms of Big Data. Artificial load tests can be done to simulate the future performance bottlenecks.*

What is Big Data?

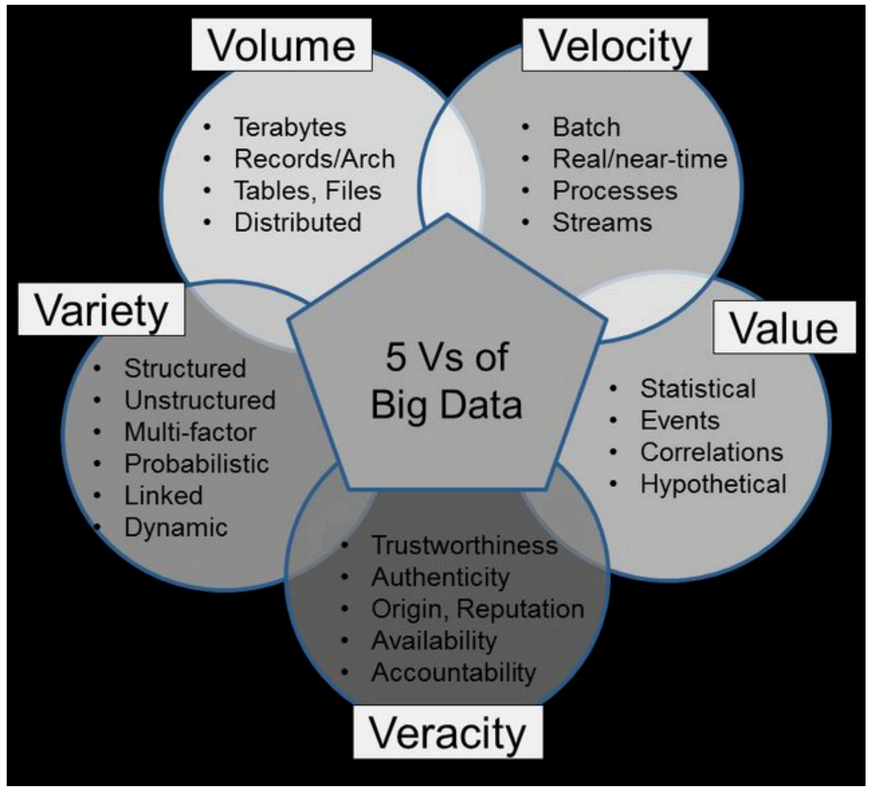
Source: (2015). A Brief Introduction on Big Data 5Vs Characteristics and Hadoop Technology. Procedia Computer Science, 48, 319-324. DOI: 10.1016/j.procs.2015.04.188

- Big Data?
  - - Big Data is something so huge and complex that is impossible for traditional systems and traditional data-warehousing tools to process and work on them.
- Big Data analytics?
  - - A big data analytics is the process if examining large amounts of data
- The challenges include:
  - - Capturing, analysis, storage, searching, sharing, visualization, transferring, and privacy

5Vs of Big Data

(2015). A Brief Introduction on Big Data 5Vs Characteristics and Hadoop Technology. Procedia Computer Science, 48, 319-324. DOI: 10.1016/j.procs.2015.04.188

- **Volume:** The benefit gained from the ability to process large amounts of information is the main attraction of big data analytics.
- **Velocity:** Refers to the speed at which new data is generated and at which data moves around.
- **Variety:** It is not always to put big data into relational database. Dealing with a variety of structured and unstructured data greatly increases the complexity of both storing and analyzing Big Data. 90% of data generated is unstructured form.
- **Veracity:** The quality of the data being captured can vary greatly. The data accuracy of the analysis depends on the veracity of the source data.
- **Value:** It is all well and good having access to big data but unless we can turn it into value it is become useless. It becomes very costly to implement IT infrastructure systems to store big data, and business are going to require a return on investment.



Source: https://www.researchgate.net/publication/273945634_Big_Security_for_Big_Data_Addressing_Security_Challenges_for_the_Big_Data_Infrastructure/figures?lo=1

Handling Big Data in RDBMS

Reference paper: *Xie, Q., Yang, W., & Yao, L. (2019). A Database Optimization Strategy for Massive Data Based Information System. In Proceedings of the 2nd International Conference on Mathematics, Modeling and Simulation Technologies and Applications (MMSTA 2019). DOI: 10.2991/mmsta-19.2019.47*

- Paper proposes single table optimization methods for massive amount of data.
- Test environment: 200 G data over 80.000.000 rows.
- Emphasize the **physical design phase of the database.**

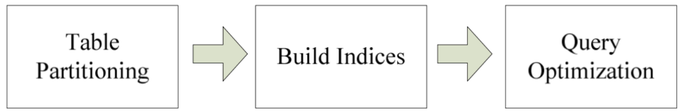


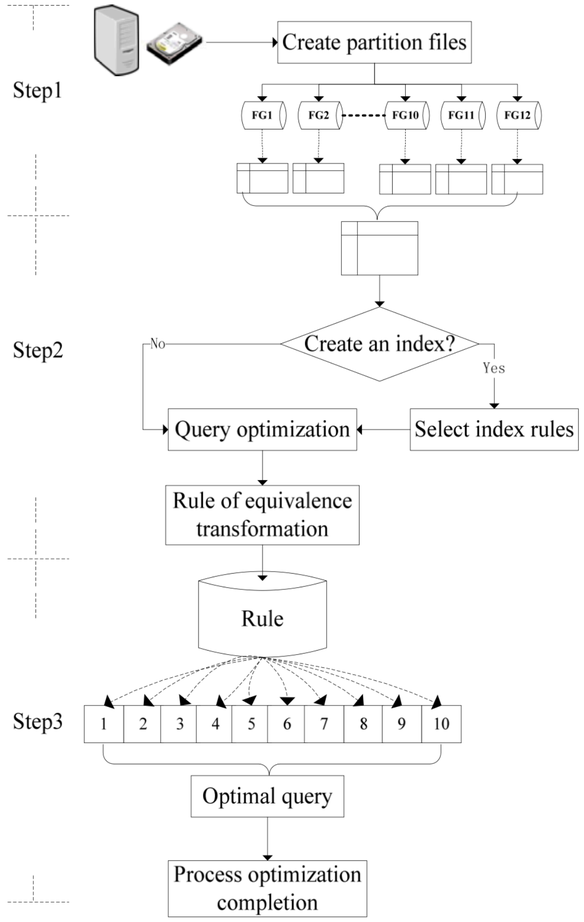
FIGURE I. THE GENERAL BLOCK OF OPTIMIZING PROCESS.

- 1) **Table Partitioning:** *Table partitioning refers to partitioning a large table physically into multiple small table storages in a database*
- 2) **Indexing:** *A quick query of data tables.*
- 3) **Optimization of the query:** *Generating the optimal SQL query statement through the equivalent transformation rules of relational algebra.*

.

Handling Big Data in RDBMS

Reference paper: *Xie, Q., Yang, W., & Yao, L. (2019). A Database Optimization Strategy for Massive Data Based Information System. In Proceedings of the 2nd International Conference on Mathematics, Modeling and Simulation Technologies and Applications (MMSTA 2019). DOI: 10.2991/mmsta-19.2019.47*



Why RDBMS is not proper for Big Data?

- Will be filled with the disadvantages of using RDMS especially for Big Data.

Advantage of switching from RDBMS to NoSQL for Big Data

- Will be filled with the advantages switching from RDMS to NoSQL for Big Data.

Handling Big Data in NoSQL

- Will be filled with the strategies to handle Big Data in NoSQL

Performance optimization of PosgreSQL for Textual Data

- The textual data that the company(that I am working for) have in PosgreSQL will be used to conduct experiments.

Experiments

- Results of the performance optimization will be mentioned.

Conclusion

- The general conclusion will be summarized.

Future work

- What can be done more especially in terms of master thesis

References

- The reference papers that are used to prepare this slide.