RESEARCH ARTICLE OPEN ACCESS

Big Data Analytics& Cloud Computing: Google BigQuery vs. Spark SQL

[1],Umi Nur Inayah[2], Fahrid Idrisa[3], Diwi Apriana [4] Kaseri
Master of Computer Science, Faculty of Information and Technology Budi Luhur University, Jakarta, Indonesia
E-mail : {[1]1711600930, [2]1711600195, [3]1711601078, [4]1711601086}@student.budiluhur.ac.id

Abstract:

Entering the era of "Big Data" development, so much data management with scalability is bad, where many moving companies experience limitations in handling large amounts of data so that is creates a variety of problems, ranging from installation and maintance difficultes, therefore it is deemed neccesary for companies to use big data by utilizing could computing, because they complement each other. In this study the researcher used the K-Nearest Neighbors and Naive Bayes algiorithm with Google BigQuery and Spark SQL database management relations systems to discuss diagnoses in patient who are likly to suffer from asthma. The results show that have two algorithms namely K-Nearest Neighbors and Naive Bayes produce relatively the same answer even though the output values are different. Algorithm if with K-Nearest Neighbors look for the nearest K value to diagnose asthma while naive bayes uses probability in the form of a percentage to ditermine the diagnoses of asthma. Then, when measuring data processing using Google BigQuery it everages 0.019 persecond for 13 rows and on Spark SQLfor an average of 0.065 persecond. So it can be concluded that measurements in the processing of asthma data in seconds result in Google BigQuery faster then Spark SQL.

Keywords—Big Data Analytics, Cloud Computing, Google BigQuery, K-Nearest Neighbor Algorithm, Naive Bayes Algorithm, Relationional Database Management System.

I. INTRODUCTION

Nowdays, a lot of data storange is starting to be stored in the *could* where the data is increasingly and continuously increasing into a lot of data, known as *Big Data (BD)*. The data is fairly complex so it needs to be stored, processed and well organized so that in the future it can be utilized [1] to obtain relevant information. The data that is too large so not possible to bestored on an ordinary server, therefore the existence of Cloud *Computing (CC)* will ease the burden on humans who may have difficulties in processin data individually without the help of computer machines. Especially if the data that needs to be processed must be done in a short time.of course to manage this large data will not be able to be done manually.this researah is related to *Big Data Analytics (BDA*),which acts as a large –scale data provider and has a sophisticated and measurable computing infrastructure[3].in this paper,we will analyze and compare the *Relational Data Base Management System (RDBMS),*which is Google BigQuery with Spark SQL. Where both of them function to collect large data in a short period of time with super-fast queries. Spark SQL it self is a module developed by Apache Spark which functions to help processing data on a large scale usually used for calling SQL,multimedia data and data streams [4]. The Google Cloud Platform releases the same product as Spark, Google BigQuery is one of the CC solutions that is useful for data analysis, data science and big data [5]. Comparison of these two RDBSM, for the classification of our data using K-Nearest Neighbor (K-NN) algorithm and Naive Bayes Algorithm.

II. Literature Review

The importance of managing BD well is very necessary for companies that have data with high

levels of complexity. Until now many solutions have been to overcome BD. It can be seen from the many applications and alternative technologies that have emerged, one of which is the Google BigQuery RDBMS and Spark SQL which will be discussed in this paper.

A. Big Data Analytics (BDA)

The existence of BDA, data of large scale can be collected organized and analyzed which in the future is usefull to see the flow and get information than can be utilized. For large companies in the current digital era, to make the right decision,it is necessary to first analyze it by utilizing BDA. Thus the company will be right on target in making decisions because it has been provided with the results of analysis of previous data processing. BDA is not only applied to certain fields, say in the field of information technology. But this But this BDA is unlimited, meaning it can be used also in other industrial fields. Because BDA has the ability to develop plat forms and new tools to store, process and speed up data in large capacity [6].

B. Cloud Computing (CC)

Basically simple cloud computing (CC) is a technology that makes the internet the foundation. Thus to use CC, it is necessary to connect to the internet first, and then users can use applications that are used to manage data. Google Drive is an example of CC implementation develoed by google. Other examples such as Dropbox Blue Cord Initiative from IBM and many more. According to [7] CC is computing requires the interne to access all sevices which will be used simultaneously more than one user.

C. Google BigQuery

Google BigQuery is referred to as the right tool for BD analysis and is capable of reaching terabytes and petabytes and processing data very quickly in just second [8]. According to [9], Goggle BigQuery is the most appropriate solution for big data storage in the cloud. Google BigQuery offers a powerful REST-API for streaming data into tables or loading it via *batch-job[10].*

D. Spark SQL

In BD processing, hadoop MapReduce developed the apache spark platform which will help increase the acceleration of processing data at large capacity [11]. Spark is an algorithm designed to process memory storage support, as well as efficient error storage recovery, where spark consists of two APIs, which consist of frames, spark SQL DataSet frame is one interface that must be used by most users. According to [12] apache Spark is processing data in a large scale memory with the addition of various master and slave nodes [13].

E. K-Nearest Neighbors (K-NN)

The algorithm is reffered to as one of the classification methods by classying it by looking at the similarity of object with *class* {14}. K-NN is the most widelyused data for data meaning because it is simple and the results are accurate [15]. Whereas accoding to [16] K-NN is a method that uses the *training*phase that is, this algorithm only stores feature *vectors* and classification of *sample training* data.

F. Naive Bayes

Naive Bayes comes from the Bayes rule, which is a simple method by looking at existing conditions and opportunities or probabilities in each condition [17]. Naive Boyes is probability calculation method. Naive Bayes is uses the Bayes theorem, the theorem in statistics to find opportunities from one class from each group of attributes that exist and determine the class where the most optimal.

III. Research Methods

The researcher used the K-NN and Naive Bayes algorithms with Google BigQuery RDBMS and Spark SQL in processing data to detemine asthma.

A. Google BigQuery vs. Spark SQL

In table 3.1 shows the differences and similarities of google bigquery and spark sql in terms of specifications and provenance.

Information	Google BigQuery	Spark SQL
Definition	Data wrehouse services on a large scale only by adding tables	Component above 'Spark Core' for structured data processing.
Primary database model	RDBMS	RDBMS
Secondary database models	Key value store	Key value store
Developer	Google	Apache software foundation
Year of release	2010	2014
Licence	Paid	free
Supported programming languages	.Net Java Java script Objective-C PHP Python Ruby	Java Python R Scala

Figure 3.1 comparison of Google BigQuery vs. Spark SQL

B. *K-Nearest Neighbors (K-NN)*

According to [18] K-NN is a valid data classification technique. K-NN thats *supervised learning* algorithm method, where most classes appear to be the result of classification. The K-NN formula [19] there :

$$D(x,y) = \sqrt{\sum_{k}^{n} = 1 \, (x_{testing} + y_{testing})^2}$$

information :

$x_{training}$: Data training ke-i
$y_{testing}$: Data testing
i : Record (line) to-i from table
n : amount of training data

In the research conducted by [20], explain the steps of K-NN, there are :

1. Determine the parameter K (value of nearest neighbor)
2. Calculate the euclidean distance of each object against the sample data
3. Sort each object into groups that have the smallest euclidean distance
4. Collect Y category (Nearest Neighbor Classification)

By using the mojority category, the classification result are obtained.

C. *Naive Bayes*

Naive Bayes in this study, the researcher made a decision by knowing the probability or oppurtunity of patients susppected of having asthma. When viewed in a formula, it is as follows :

$$p(A|B) = \frac{p(B|A)+p(A)}{p(B)}$$

Information :

- p(A|B) is probability A due to B
- p(B|A) is probability B due A
- p(A) is probability A regardles of any factor
- p(B) is probability B regardles of other factors.

IV. Result and Discussion

In this study, the researcher will compare two Google BigQuery RDBMS with spark SQL and for their case studies classify asthma diagnoses using two classifications of the K-NN algorithm and Naive Bayes. First, the researcher prepare questionnaire as in Table 4.1 for asthma patients.

Code	Questions	Yes/no
G1	Are there breath sounds (wheezing)?	
G2	Is cough	
G3	Whether shortness of breath suddenly?	
G4	Is intensity of severe tightness?	
G5	Is the chest feel heavy?	
G6	Wheather restless?	
G7	Whether shortness of breath relapse?	
G8	Is the intensity of crowded from low to medium?	
G9	Is sometimes there feel a breath (wheezed) or not?	
G10	Whether Sometimes cough?	
G11	Whether shorthness of breath often relapse because of dust, the smoke, and the cold air?	

Figure 4.1 asthma questionnaire

A. *Disease Diagnosis*

The researcher make 3 categories of astma diagnoses, namely chronic asthma, and periodic asthma. Table 4.2 shows the names of asthma based on the names of asthma based on the diagnosis experienced by the patient.

Asthma	Cough	Shorthness of breath	diagnoses
V	V	V	Asthma chronic
		V	Asthma Acute
	V	V	Asthma periodic
V			Asthma Acute
V		V	Asthma chronic
V	V		Asthma periodic

Table 4.2 Diagnosis of Asthma

B. *K-NN And Naive Bayes*

Figure 4.1 in the form of a diagnosis of chronic asthma using the K-NN algorithm:

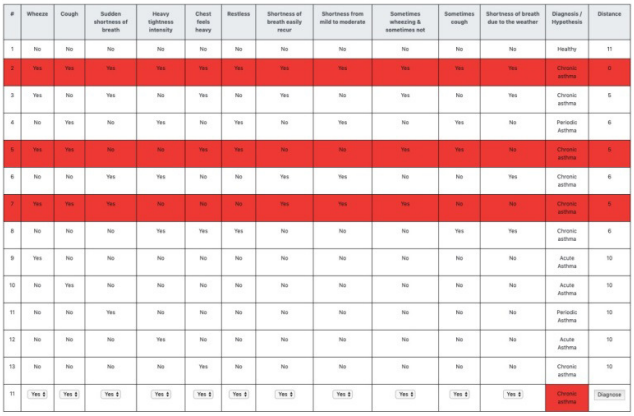

Figure 4.1 Diagnosis of Asthma with K-NN

Based on figure 4.1 displays the results for the diagnosis of chronic asthma with known value k=3. Patients answer all questions as in table 4.1 with answers yes all means the closest answer is cronic Asthma with a diagnostic value = 0. Then because the value of K is filled = 3, then there are 2 more diagnoses, as in figure 4.1, the second K value = 5 and the third K value is the same ”5”, but with a different disease diagnosis.

Next in figure 4.2 is the result of diagnosing the disease using Naive Bayes. Patients fiull out questions by answering yes to coughing and tightness, not answer to wheezing. Produce the most presentations to the diagnosis of periodic asma as much as 72%.

The Name of the disease	Asthma Acute	Asthma chronic	Asthma periodic
Information	9%	19%	72%

Figure 4.2 Diagnosis of Asthma with Naive Bayes

When carrying out data retrieval with a database from Google BigQuery the average timeit takesis around 0.019 seconds to display 13 data lines. Whereas if you use spark SQL, The average time produced is around 0.065 seconds.

V. Conclusion

Based on the research that has been done then the researcher conclude some conclusions as follows :

:

1. The process of testing the diagnosis of asthma using K-Nearest Neighbors and Naive Bayes the differences in the calculation method and the output results.
2. Measurements in the processing asthma data in second produce Google BigQuery faster than Spark SQL.

DAFTAR PUSTAKA

[1] S. Khan, K. A. Shakil, and M. Alam, “Cloud-Based Big Data Analytics - A Survey of Current Research and Future,” *AISC - Big Data Anal.*, vol. 654, no. 4, pp. 595–604, 2017.

[2] J. L. Reyes-Ortiz, L. Oneto, and D. Anguita, “Big Data Analytics in the Cloud: Spark on Hadoop vs MPI/OpenMP on Beowulf,” *Procedia Comput. Sci.*, vol. 53, no. 1, pp. 121–130, 2015.

[3] C. Yang, Q. Huang, Z. Li, K. Liu, and F. Hu, “Big Data and cloud computing: innovation opportunities and challenges,” *Int. J. Digit. Earth*, vol. 10, no. 1, pp. 13–53, 2017.

[4] M. Al-Zobbi, S. Shahrestani, and C. Ruan, “Experimenting sensitivity□based anonymization framework in apache spark,” *J. Big Data*, vol. 5, no. 1, p. 38, 2018.

[5] K. Yue, “Querying Bitcoin Blockchain Using

SQL," pp. 1–17, 2018.

[6] H. Akhavan-Hejazi and H. Mohsenian-Rad, "Power systems big data analytics: An assessment of paradigm shift barriers and prospects," *Energy Reports*, vol. 4, no. 2, pp. 91–100, 2018.

[7] A. R. Hakim, "Analisis Perbandingan Sistem Cloud Azure Dan Google Cloud," *InfoTekjar-Jurnal Nas. Inform. dan Teknol. Jar.*, no. 9, pp. 38–41, 2016.

[8] D. Fernando, "Visualisasi Data Menggunakan Google Data Studio," *Semin. Nas. Rekayasa Teknol. Informasi(SNARTISI)*, no. November, 2018.

[9] S. Khan, K. A. Shakil, and M. Alam, "PABED – A Tool for Big Education Data Analysis Samiya," *Comput. Soc.*, vol. 1, no. 334, pp. 1–5, 2018.

[10] R. Feist, "Cloudbased Production Optimization - Potential and Limits Today," *Appl. Mech. Mater.*, vol. 885, pp. 48–55, 2018.

[11] S. Oliviandi, A. B. Osmond, and R. Latuconsina, "Implementasi Apache Spark Pada Big Data Berbasis Hadoop Distributed File System," *e-Proceeding Eng.*, vol. 5, no. 1 Maret, pp. 1005–1012, 2018.

[12] S. M. Othman, F. M. Ba-Alwi, N. T. Alsohybe, and A. Y. Al-Hashida, "Intrusion detection model using machine learning algorithm on Big Data environment," *J. Big Data*, vol. 5, no. 24 September, 2018.

[13] N. Mahasivam, N. Nikolov, D. Sukhobok, and D. Roman, "Data preparation as a service based on Apache Spark," *Lect. Notes Comput. Sci.*, vol. 10465, no. 01 September, pp. 125–139, 2017.

[14] Y. I. Claudy, R. S. Perdana, and M. A. Fauzi, "Klasifikasi Dokumen Twitter Untuk Mengetahui Karakter Calon Karyawan Menggunakan Algoritme K-Nearest Neighbor (KNN)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 8, pp. 2761–2765, 2018.

[15] I. Triguero, J. Maillo, J. Luengo, S. Garcia, and F. Herrera, "From Big data to Smart Data with the K-Nearest Neighbours," *IEEE Int. Conf. Internet Things IEEE Green Comput. Commun. IEEE Cyber, Phys. Soc. Comput. IEEE Smart Data*, pp. 859–864, 2016.

[16] B. P. Nugroho, "Implementasi sistem untuk prediksi harga emas," vol. 8, no. 2, pp. 90–104, 2018.

[17] S. Fanissa, M. A. Fauzi, and S. Adinugroho, "Analisis Sentimen Pariwisata di Kota Malang Menggunakan Metode Naive Bayes dan Seleksi Fitur Query Expansion Ranking," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 8, pp. 2766–2770, 2018.

[18] A. Y. Saputra and Y. Primadasa, "Penerapan Teknik Klasifikasi Untuk Prediksi Kelulusan Mahasiswa Menggunakan Algoritma K-Nearest Neighbour," *Techno.COM*, vol. 17, no. 4, pp. 395–403, 2018.

[19] Mustakim,Giantika O, "Algoritma K-Nearest Neighbor Classification Sebagai Sistem Prediksi Predikat Prestasi Mahasiswa," *J. Sains dan Teknol. Ind.*, vol. 13, no. 2, pp. 195–202, 2016.

[20] M. S. Mustafa and I. W. Simpen, "Perancangan Aplikasi Prediksi Kelulusan Tepat Waktu Bagi Mahasiswa Baru Dengan Teknik Data Mining (Studi Kasus: Data Akademik Mahasiswa STMIK Dipanegara Makassar)," *Citec J.*, vol. 1, no. 4, pp. 270–281, 2014.