PEMBANGUNAN INFRASTURKTUR BIG DATA BERBASIS HADOOP PADA UNIVERSITAS JAMBI

Maryetta Hana[1)], Jefri Marzal [2)], Mauladi [3)]
Program Studi Sistem Informasi, Fakultas Sains dan Teknologi, Universitas Jambi
email: hannamaryetta@gmail.com

Abstract

Rapid and diverse data growth is a challenge for managing data. Big Data is a technology that enables data management with 3V characteristics (Volume, Variety, Velocity). This study aims to document the process of building a large data infrastructure using the method of PPDIOO. PPDIOO can be easily described as a method of analyzing the development of computer network installations developed by Cisco. Through a series of steps performed using the PPDIOO method, big data systems can be run that are tested by applying the K-Means algorithm method and Reuters-21578 data. Clustering results show that the more nodes are used the clustering process will be faster.

Keywords: *Big Data, PPDIOO, K-Means, clustering.*

1. PENDAHULUAN

Data merupakan kumpulan fakta bagi sebuah organisasi baik pemerintah maupun swasta, data dapat berupa angka, alphabet, simbol, karakter, gambar, suara, video, dan lain-lain. Data terbagi menjadi data terstruktur, data semiterstruktur dan data tidak terstruktur. Data tidak terstruktur berisi obyek atau dokumen baik ukuran maupun tipe yang bersifat bebas, sedangkan data terstruktur setiap elemen data harus mengacu pada format yang telah ditetapkan. Data pada sebuah organisasi merupakan aset yang penting, dengan adanya data organisasi dapat memiliki sebuah informasi untuk melakukan inovasi. Organisasi membutuhkan penyusunan data yang baik agar dapat memperoleh informasi yang diinginkan dengan efektif sehingga dapat memberikan informasi yang dibutuhkan oleh pemimpin ataupun *stakeholder* dalam mengambil sebuah keputusan. Data yang baik dapat disusun dalam sebuah *database*.

Database merupakan sebuah wadah untuk menyimpan data-data pada sebuah organisasi. Menurut Connolly dan Begg, basis data adalah sekumpulan koleksi data yang berhubungan secara logikal, dan sebuah deskripsi dari data tersebut, didesain untuk menemukan keperluan informasi pada sebuah perusahaan. *Database* penting bagi sebuah organisasi karena dengan adanya *database* sebuah organisasi dapat mengumpulkan, mengorganisir, dan menganalisa data.

Universitas Jambi (UNJA) merupakan sebuah organisasi pemerintah yang bergerak dalam dunia pendidikan. Teknologi informasi telah digunakan berbagai aktivitas utama pada UNJA, dimulai dari proses registrasi mahasiswa baru, kontrak perkuliahan setiap semester, pembayaran, absensi, berita acara perkuliahan, hingga hasil akademik mahasiswa setiap semester. Data lainnya mengenai penelitian, pengabdian masyarakat, absensi pegawai, dan kegiatan penunjang lainnya juga sudah dilakukan dengan menggunakan teknologi informasi.

UNJA menggunakan teknologi *Database Management System* (DBMS) MySQL dalam menyimpan data, dimana untuk melakukan analisa data dilakukan menggunakan *query*. DBMS merupakan sebuah teknologi *database* yang menyimpan data secara terstruktur dan masif, dengan menggunakan DBMS UNJA dapat mengurangi terjadinya redudansi atau duplikasi pada penyimpanan data dengan kondisi database yang di normalisasi. DBMS memiliki kekurangan berhadapan dengan *unstructured data*, dimana DBMS hanya dapat menyimpan data dalam bentuk *text*.

Pada masa yang akan datang, dilihat dari kondisi dan pengembangan sistem yang saat ini sedang dilakukan oleh UNJA, UNJA berpotensi memiliki *unstructured data*, ketika UNJA memiliki *unstructured data* maka tidak memungkinkan tetap menggunakan teknologi DBMS MySQL untuk melakukan analisa data.

Berawal dari keberhasilan perusahaan-perusahaan *web service* raksasa seperti halnya Google dan Facebook dalam mengelola dan memanfaatkan *unstructured data* dalam volume yang sangat besar, sebuah konsep yang dikenal dengan istilah *Big Data* kemudian menjadi pusat perhatian dalam dunia teknologi informasi (Wijaya, 2015). *Big Data* jika didefinisikan secara sederhana adalah sebagai sekumpulan data dengan *volume* yang sangat besar yang terlalu kompleks untuk dapat diproses dengan teknologi pengolahan data konvensional. IBM mendeskripsikan *Big Data* sebagai suatu teknologi yang dapat melakukan pengolahan, penyimpanan dan analisis data yang sangat kompleks dalam beragam bentuk atau format (*Variety*), berukuran besar (*Volume*) dan pertambahan data yang sangat cepat (*Velocity*) yang kemudian akan dianalisa atau diolah lagi untuk keperluan tertentu seperti membuat keputusan (*decision making*), prediksi, dan lainnya.

Untuk menjawab tantangan kompleksitas pemberdayaan *Big Data*, *Apache Software Foundation* (ASF) menciptakan *Apache Hadoop*, yang merupakan sebuah *framework Distributed System* yang dikembangkan melalui proyek *Open Source* (White, 2012). Apache Hadoop ini terdiri atas Hadoop Distributed File System (HDFS) dan Hadoop MapReduce. Dalam hal ini HDFS bertugas sebagai sistem penyimpanan data secara terdistribusi atau bertugas untuk sinkronisasi data, sedangkan Hadoop MapReduce berfungsi sebagai *framework* pemrosesan data secara paralel dan terdistribusi. Shvachko et al. (2012) dalam jurnalnya menjelaskan secara rinci tentang pengalamannya menggunakan Hadoop dalam mengelola sekitar 25 petabyte data perusahaan di Yahoo!. Data-data tersebut disimpan dan diproses dalam sejumlah *cluster* yang secara total terdiri atas 25.000 komputer *server*. Diantara *cluster-cluster* komputer tersebut, *cluster* yang paling besar terdiri atas 3.500 komputer server. Mereka menemukan bahwa dalam hal *scalability* Hadoop memang memiliki keunggulan yang tak dapat dipungkiri. Selain itu mereka juga menyarankan bahwa mengelola Hadoop dalam beberapa *cluster* yang lebih kecil akan lebih menguntungkan daripada mengelola Hadoop dalam satu *cluster* yang besar.

Dari penjelasan tersebut teknologi *big data* merupakan solusi yang tepat untuk menjawab tantangan kompleksitas analisis data pada UNJA. Pernyataan Asniar dalam jurnalnya menyimpulkan bahwa ledakan data di perguruan tinggi memungkinkan untuk memanfaatkan *big data analytic* berupa *learning analytic*, *academic analytic*, dan *process analytic*.

Tujuan dilakukannya penelitian ini adalah untuk memahami teknik penerapan teknologi *big data* berbasis Hadoop, mengimplementasikan teknologi, dan pola yang tepat untuk membangun infrastruktur *big data*.

2. METODE PENELITIAN

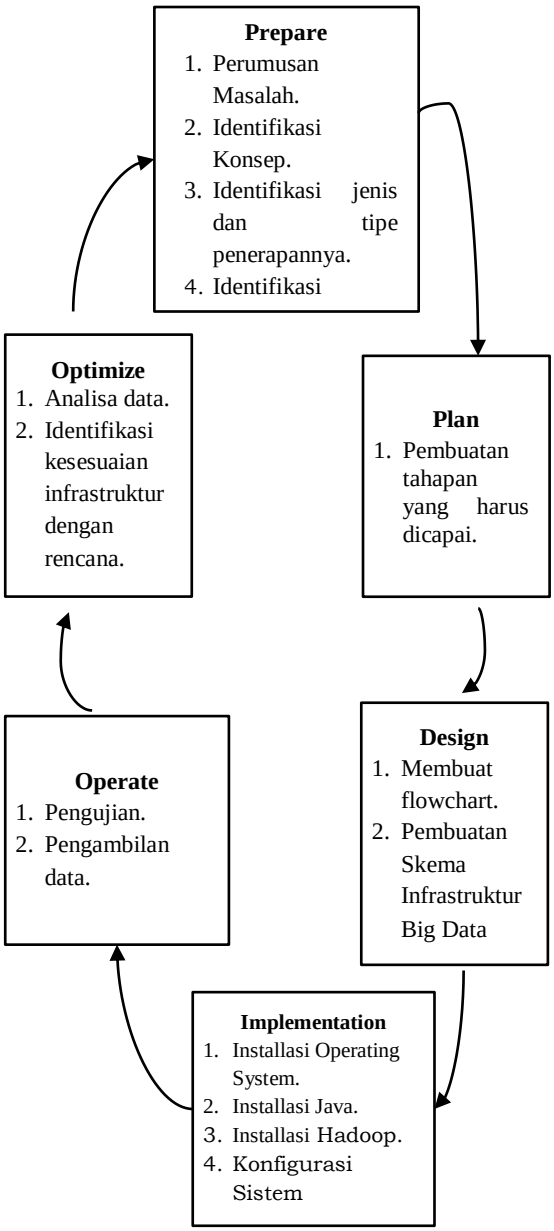
Bahan penelitian yang digunakan bersumber dari bank data UCI Machine Learning *Reuters-21578 Text Categorization Collection Data Set.* Perangkat keras yang digunakan adalah Laptop atau Notebook dengan spesifikasi RAM min. 4 GB, processor intel i5 dengan 2.20 GHz dan kapasitas penyimpanan data yang free min. 200 GB.

Metode PPDIOO

Penelitian ini merupakan penelitian eksperimen, metode PPDIOO digunakan untuk membangun sistem Big Data. Metode PPDIOO merupakan metode analisis hingga pengembangan instalasi jaringan komputer yang di kembangkan oleh Cisco pada materi Designing for Cisco Internetwork Solutions (DESGN) yang mendefinisikan secara terus menerus siklus hidup layanan yang dibutuhkan untuk pengembangan jaringan komputer.

a. Analisis kebutuhan infrastruktur
   Pada tahapan ini dianalisis kebutuhan *hardware* dan *software* yang digunakan untuk membangun infrastruktur *big data*.
   1. Sumber data atau dataset dari bank data UCI Machine Learning mengenai *Reuters-21578 Text Categorization Collection Data Set.*

Daftar kebutuhan *hardware* diperoleh dengan melakukan studi pustaka mengenai kebutuhan yang diperlukan untuk membangun infrastruktur .



3. HASIL DAN PEMBAHASAN

a. Tahap Prepare

Analisis kebutuhan infrastruktur.
Pada tahapan ini dianalisis kebutuhan *hardware* dan *software* yang digunakan untuk membangun infrastruktur *big data*.

1. Sumber data atau dataset yang digunakan dari bank data UCI Machine Learning mengenai *Reuters-21578 Text Categorization Collection Data Set.*
2. Daftar kebutuhan *hardware* diperoleh dengan melakukan studi pustaka mengenai kebutuhan yang diperlukan untuk membangun infrastruktur *big data*. Dari hasil studi pustaka yang dilakukan didapatkan beberapa *requirements* yang disajikan dalam tabel 1.

Tabel 1. Kebutuhan Hardware

Nama *Hardware*	Spesifikasi *Hardware*	Fungsi
Personal Computer	1. Ruang *Harddisk* tersisa min. 40GB 2. Memory min. 2GB 3. Processor min. Intel Core i3 atau yang setara.	Personal Computer atau PC ini akan digunakan sebagai media untuk mengkonfigurasi sistem *big data* yang akan dibangun. Jumlah PC yang akan digunakan disesuaikan dengan kebutuhan *cluster*.
Router / Switch	Optional	Router digunakna sebagai media untuk menghubungkan antar PC agak dapat saling berkomunikasi.
RJ45	-	RJ45 atau kabel RJ45 digunakan sebagai media yang menghubungkan PC dengan Router/Switch agar

	dapat saling berkomunikasi.

3. Daftar kebutuhan *software* diperoleh dengan melakukan studi pustaka mengenai kebutuhan yang diperlukan untuk membangun infrastruktur *big data*. Dari hasil studi pustaka yang dilakukan didapatkan beberapa *requirements* sebagai berikut:
   a. OS Linux Ubuntu 16.04 LTS
   b. Browser (Google Chrome, Mozilla Firefox, dll)
   c. Java (openjdk 8)
   d. Hadoop 2.7.2
   e. SSH
   f. Maven
   g. Mahout

Dari penjelasan spesifikasi yang dipaparkan sebelumnya terdapat keterbatasan dalam implementasinya diantaranya adalah keterbatasan dalam hal alat untuk menerapkan spesifikasi analisa di awal sehingga dalam penelitian ini menyesuaikan spesifikasi tersebut untuk dapat membuat *prototype* dari infrastruktur *big data*. Spesifikasi yang digunakan disajikan dalam tabel 2.

Tabel 2. Spesifikasi Komputer Cluster

Jenis Node	Memory	Hard Drive
Master	800MB	40GB
Slave 1st Node	800MB	40GB
Slave 2nd Node	800MB	40GB

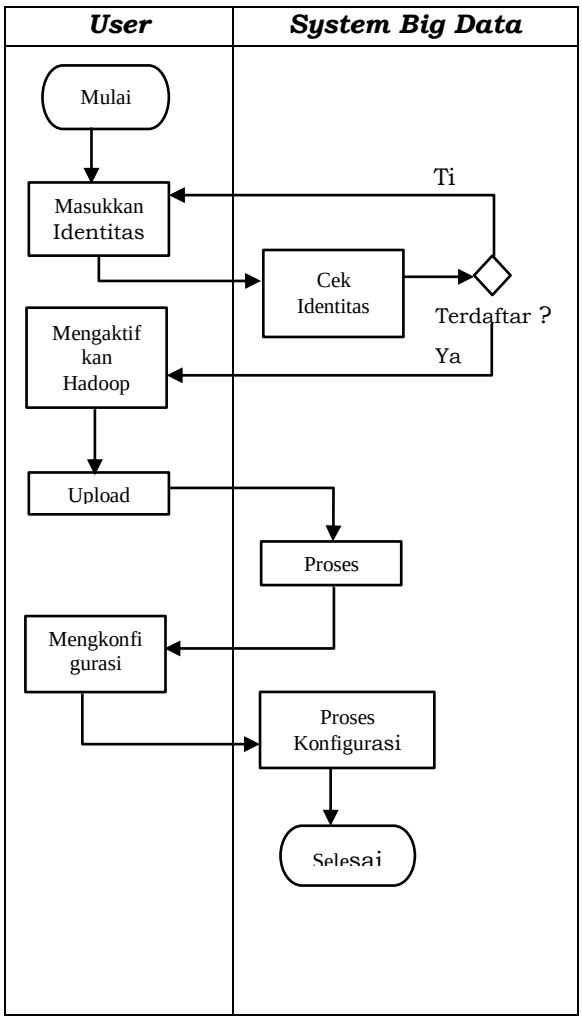
b. Tahap Plan

Berdasarkan proses implementasinya, sistem *big data* pada mulanya dikembangkan dalam masing-masing *node*. Implemetasi sistem *big data* pada sebuah komputer disebut *single node cluster*. Setelah sistem big data diimplementasikan pada masing-masing *node*, maka langkah selanjutnya adalah menyatukan seluruh *node* menjadi satu, proses ini dinamakan *multi node cluster*.

Tahap Design

Terdapat 2 (dua) desain yang dibuat yaitu *flow chart diagram*, dan skema sistem *big data*.

a. *Flow chart diagram*



Gambar 1. Flow chart Diagram

User melakukan *login* dengan cara memasukkan data yang berupa *username* dan *password* kedalam sistem dan jika terdaftar maka akan diberikan izin untuk masuk kedalam sistem dan melakukan konfigurasi. Jika *user* salah mengisikan data *login* maka sistem akan meminta untuk mengulang memasukkan data. Setelah berhasil *login* kedalam sistem maka *user* akan melakukan konfigurasi awal dengan mengaktifkan hadoop (start-all.sh). Dilanjutkan oleh user untuk meng-*upload* data yang akan diproses oleh sistem. Maka user melakukan konfigurasi output yang dinginkan, setalah konfigurasi dilakukan maka sistem akan

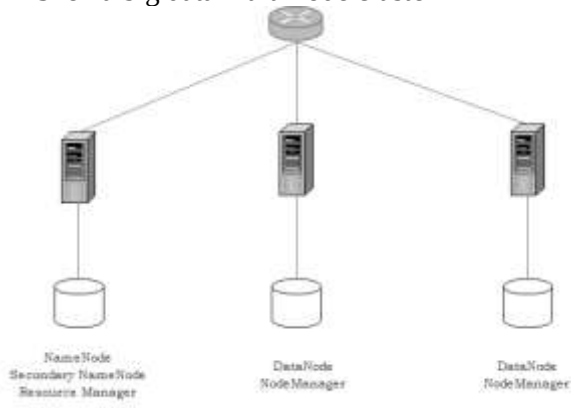
memproses konfigurasi tersebut dan akan menampilkan hasil dari konfigurasi yang telah dibuat.

b. Skema *big data single node cluster*

Gambar 2. Skema single node cluster

Skema *single node cluster* yang memperlihatkan sistem Hadoop yang hanya diimplementasikan pada sebuah komputer dapat dilihat pada gambar 16. *Single Node* (komputer tunggal) tersebut bertindak sebagai *master node* dan *slave node*. Sehingga *service* yang dijalankan pada *node* ini adalah *NameNode, Secondary NameNode, NodeManaer, Recource Manager, DataNode*. *Router* digunakan untuk membuat koneksi jaringan lokal. Sehingga setiap komputer tunggal memiliki IP Address. Adapun setiap IP Address yang dimiliki komputer dapat saling terhubung kedalam satu jaringan. Sehingga, ketika setiap komputer memliki IP Address yang saling terhubung dalam satu jaringan dapat dikonfigurasi ke dalam *multi node cluster*.

c. Skema big data multi node cluster



Gambar 3. Skema multi node cluster

Skema *multi node cluster* yang memperlihatkan sistem Hadoop yang diimplementasikan pada multi komputer dapat dilihat pada gambar 17. Perancangan *multi node cluster* menggunakan 3 komputer yang terdiri dari 1 *master node* dan 2 *slave node*. *Router* digunakan untuk membuat koneksi jaringan lokal. *Router* bertindak sebagai *gateway* pada jaringan yang berfungsi untuk menghubungkan antar *node*. IP Address digunakan untuk memberikan alamat pada masing-masing *node*. *Master node* akan menjalankan *service Namenode, Secondary NameNode,* dan *Resource Manager*. Sedangkan *slave node* akan menjalankan *service DataNode* dan *NodeManager*.

Tahap Implement

Perancangan sistem *big data* dimulai dengan melakukan konfigurasi dari masing masing komputer, tiap *virtual machine* di i*nstall* dengan *operating system* Linux Ubuntu 16.04 LTS, setelah itu proses dilanjutkan dengan menggabungkan semua *node* menjadi satu.

Tahap Operate

Hasil yang diperoleh dari algoritma K-Means adalah berupa direktori kmeans (hadoop-mahout/kmeans) pada HDFS. Mahout menyediakan metode untuk menganalisa hasil dari komputasi K-Means. Metode yang digunakan adalah dengan menggunakan perintah clusterdump. Metode ini dapat membuat file analisa yang mengelompokkan item data berdasarkan centroid atau pusat kelompoknya. Hasil dari clusterdump mengasilkan sebanyak 20 cluster disajikan dalam tabel 3.

Tabel 3. Hasil clusterdump

VL-2115, n = 348	VL-213, n = 900
VL-12337, n = 825	VL-1142, n = 954
VL-1584, n = 760	VL-9182, n = 911
VL-10838, n =1416	VL-13579, n = 1072
VL-8645, n = 2327	VL-17541, n = 621
VL-2455, n = 2395	VL-19501, n = 442
VL-12924, n = 1061	VL-11181, n = 1062
VL-11166, n = 559	VL-11520, n = 2541
VL-14856, n = 775	VL-19350, n = 940
VL-13846, n =1371	VL-5298, n = 298

JUMLAH	**n = 21578**

VL-xxxx merupakan identitas yang secara otomatis diberikan oleh Mahout. “n” merupakan total data pada kelompok tersebut. “c” merupakan centroid atau pusat kelompok akhir, “r” merupakan radius dari kelompok, data ini tidak dapat disajikan dalam tabel.

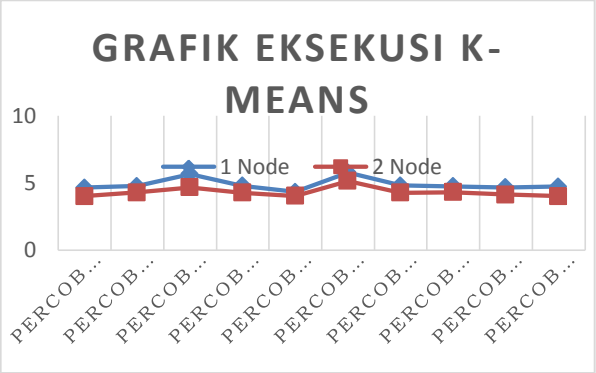
Tahap Optimize

Di tahap ini dilakukan analisa dari hasil yang telah didapatkan pada tahap operate, setelah itu dilakukan identifikasi apakah sistem yang berjalan telah sesuai dengan sistem yang direncanakan. Pada tahap-tahap sebelumnya hasil rancangan yang ada pada tahap design dapat diterapkan dengan baik pada tahap operate. Data yang didapatkan pada bank data dapat diolah oleh sistem, hal ini dapat di tunjukkan dengan samanya jumlah data sebelum di olah dengan sesudah diolah yaitu sebanyak 21.578 data.

Tabel 4. Hasil eksekusi K-Means

Percobaan ke -	Waktu eksekusi K-Means pada Slave Node (Menit)	
	1	2
1	4.64	4.02
2	4.76	4.28
3	5.65	4.66
4	4.78	4.27
5	4.32	4.04
6	5.77	5.15
7	4.81	4.26
8	4.75	4.31
9	4.66	4.13
10	4.73	4.02
Rata-rata	**4.89**	**4.31**

Pada tahap ini dilakukan pemantauan dan pengujian sistem big data dengan menjalankan komputasi K-Means menggunakan library Mahout sebanyak 10 kali dalam jumlah *slave node* yang berbeda. Nilai rata-rata digunakan untuk mengevaluasi. Tabel 4 menunjukkan bahwa semakin banyak jumlah *slave node* maka waktu eksekusi K-Means semakin cepat. Hal ini membuktikan bahwa sistem sudah berjalan dengan baik sehingga tidak diperlukan perubahan.



Gambar 4. Grafik Eksekusi K-Means

Data yang diujicobakan pada penelitian ini merupakan data klasifikasi teks yaitu Reuters-21578. Data ini merupakan data klasifikasi teks yang sering diujicobakan oleh peneliti lain untuk menguji sebuah algoritma (Lewis), dengan menggunakan data ini diasumsikan bahwa data yang akan dimiliki UNJA dapat diolah menggunakan sistem big data yang dibangun dengan menggunakan algoritma K-Means.

Berdasarkan hasil yang diperoleh, menunjukkan proses *clustering* yang dilakukan oleh algoritma K-Means dilakukan dengan benar, hal ini dapat dibuktikan dengan samanya jumlah data sebelum proses *clustering* dengan sesudah *clustering*.

Proses *clustering* dalam *big data* terkait dengan *framework* Hadoop yang dijelaskan oleh Wijaya (2015), bahwa Hadoop merupakan sistem terdistribusi *open source* yang menerapkan *programming model* sederhana terdiri dari HDFS dan MapReduce yang ditujukan untuk memproses data berukuran raksasa dalam suatu *cluster* komputer.

Hasil penerapan metode PPDIOO didapatkan bahwa sistem big data dapat diimplementasikan dengan menggunakan *framework* Hadoop, didalam framework Hadoop terdapat banyak *tools* yang dapat diinstall sesuai dengan kebutuhan instansi atau *company* yang akan membangun sistem *big data*. Penelitian ini menggunakan *tools* maven dan mahout untuk menerapkan metode algoritma K-Means. Hasil penelitian ini telah menjawab pertanyaan penelitian yang diajukan yaitu bagaimana membangun infrastruktur teknologi *big data* berbasis Hadoop. Penelitian ini membuktikan bahwa proses clustering dapat dilakukan lebih cepat ketika menggunakan lebih dari 1 node dengan

mengimplementasikan skema *multi cluster* yang ada pada tahap plan.

Dari hasil penelitian dapat direkomendasikan untuk pembangunan sistem big data pada UNJA dapat dilakukan dengan menerapkan framework Hadoop dengan menggunakan tools maven dan mahout yang berfungsi untuk menerapkan algoritma K-Means, hal yang dapat diterapkan menggunakan sistem big data dengan algoritma K-Means di UNJA adalah pemanfaatan data mahasiswa dari sejak terdaftar, proses perkuliahan, sampai selesai, jika proses perkuliahan dilakukan secara elektronik dipastikan data yang tersimpan akan sangat beragam, data tersebut dapat dikelola untuk dianalisia. Sebagai contoh analisa yang mungkin dilakukan adalah melakukan pengklasteran untuk mahasiswa pelamar beasiswa, dalam ruang lingkup akademik dalam skala yang besar, penerimaan pelamar beasiswa dapat dilakukan dengan sistem big data. Dalam sektor kesehatan pemanfaatan sistem big data dengan menggunakan algoritma K-Means juga dapat diterapkan, contoh penerapan yang memungkinkan adalah untuk menganalisa jenis penyakit yang diderita oleh pasien denga memanfaatkan sistem big data dan algortima K-Means proses pencocokan data yang dialami pasien dapat di cari lebih cepat sehingga diagnosis dari keluhan pasien hasilnya dapat lebih akurat.

4. KESIMPULAN

Infrasturktur Big Data dapat di implementasikan dengan menggunakan framework Hadoop, Maven, Mahout, dan algoritma K-Means dengan benar. Hal ini dapat ditunjukkan dengan samanya jumlah "n" pada hasil cluster dengan jumlah data sebelum di proses oleh algoritma K-Means.

Dari hasil penelitian pembangunan infrastruktur big data berbasis Hadoop ini, ada beberapa saran untuk penelitian selanjutnya dengan topik yang di paparkan sebagai berikut :

1. Menggunakan algoritma data mining yang berbeda.
2. Membuat Data Visualization agar hasil analytic dapat dipresentasikan sehingga mudah di mengerti *user*.

5. REFERENSI

Apache Software Foundation. *Apache Hadoop*. ((http://hadoop.apache.org/) 19 April 2017).

Asniar. 2015. *Penggunaan Big Data Analytic di Perguruan Tinggi*. Universitas Telkom.

Connolly, T. dan Begg, C. 2005. *Database Systems A Pratical Approach to Design, Implementation and Management Edisi 4*. Pearson Education Limited. Inggris.

Lewis, David D. 1998. Naive (bayes) at forty: The independence assumption in Information retrieval. pages 4–15. Springer Verlag.

White, Tom. 2012. *Hadoop: The Definition Guide 3rd Edition*. O'Reilly.

Wijaya, W.M. 2015. *Teknologi Big Data: Sistem Canggih di balik Google,Yahoo!, Facebook, IBM (Teori hingga Tutorial)*. Yogyakarta: Deepublish.