

**SKRIPSI**

**REDUKSI BIG DATA DENGAN ALGORITMA AGGLOMERATIVE  
CLUSTERING UNTUK SPARK**



**Matthew Ariel**

**NPM: 2015730010**

**PROGRAM STUDI TEKNIK INFORMATIKA  
FAKULTAS TEKNOLOGI INFORMASI DAN SAINS  
UNIVERSITAS KATOLIK PARAHYANGAN  
2019**



**UNDERGRADUATE THESIS**

**BIG DATA REDUCTION WITH AGGLOMERATIVE CLUSTERING  
ALGORITHM FOR SPARK**



**Matthew Ariel**

**NPM: 2015730010**

**DEPARTMENT OF INFORMATICS  
FACULTY OF INFORMATION TECHNOLOGY AND SCIENCES  
PARAHYANGAN CATHOLIC UNIVERSITY  
2019**



## ABSTRAK

*Big data* adalah istilah yang menggambarkan kumpulan data dalam jumlah yang sangat besar, baik data yang terstruktur maupun data yang tidak terstruktur. Kumpulan data tersebut menyimpan informasi yang bisa dianalisis dan diproses untuk memberikan wawasan kepada organisasi atau perusahaan. *Big data* dapat mencapai *petabyte* dan menghabiskan banyak tempat penyimpanan.

*Big data* perlu direduksi untuk menghemat tempat penyimpanan. Algoritma *Hierarchical Agglomerative Clustering* dapat digunakan untuk mereduksi data. Dengan bantuan sistem terdistribusi seperti Hadoop, proses reduksi data dapat dilakukan secara paralel dan lebih cepat. Sayangnya, teknologi Hadoop masih dapat dikatakan 'terlalu lambat' dalam melakukan proses reduksi data karena hasil sementara dari setiap tahap akan disimpan di *disk* sampai dibutuhkan kembali di tahap selanjutnya.

Untuk mempercepat proses reduksi data, Hadoop dapat digantikan dengan Spark. Spark adalah sistem terdistribusi, mirip seperti Hadoop. Tetapi, yang membedakan antara Hadoop dengan Spark adalah pada cara penyimpanan sementara saat melakukan proses reduksi data. Hadoop menggunakan *disk* sebagai tempat penyimpanan sementara, sedangkan Spark menggunakan memori sebagai tempat penyimpanan sementara. Pembacaan dan penulisan akan lebih cepat saat menggunakan memori dibandingkan dengan menggunakan *disk*, sehingga Spark akan lebih cepat dibandingkan dengan Hadoop.

Perangkat lunak dibuat untuk mengimplementasikan algoritma *Hierarchical Agglomerative Clustering* dalam Spark. Pengujian juga dilakukan dengan membandingkan waktu eksekusi algoritma *Hierarchical Agglomerative Clustering* saat diimplementasikan pada Hadoop dan saat diimplementasikan pada Spark. Waktu eksekusi dicatat untuk ukuran data 1GB, 2GB, 3GB, 5GB, 10GB, 15GB, dan 20GB.

Berdasarkan hasil pengujian, Spark memiliki waktu eksekusi yang lebih cepat dibandingkan dengan Hadoop pada jumlah partisi yang besar. Waktu eksekusi Spark menurun ketika jumlah partisi ditingkatkan, sedangkan waktu eksekusi Hadoop menurun ketika jumlah partisi ditingkatkan. Waktu eksekusi terbaik Spark masih lebih cepat dibanding waktu eksekusi terbaik Hadoop.

**Kata-kata kunci:** *Big Data*, Reduksi Data, *Hierarchical Agglomerative Clustering*, Spark, Hadoop



## **ABSTRACT**

Big data is a term that describes the large volume of data, both structured and unstructured. The data set stores information can be analyzed and processed to provide insight to organization or company. Big data can reach up to petabytes and takes a lot of storage spaces.

Big data need to be reduce to save storage space. The Hierarchical Agglomerative Clustering algorithm can be used to reduce data. With the help of distributed systems such as Hadoop, reduction process can be done in parallel with less execution time. Unfortunately, Hadoop can still be said to be 'too slow' in the process of data reduction because temporary results from each stage will be stored on the disk until it is needed again at a later stage.

To speed up the data reduction process, Hadoop can be replaced with Spark. Spark is a distributed system, similar to Hadoop. However, what distinguishes Hadoop from Spark is the way Spark temporarily store data. Hadoop uses disk as its temporary storage, while Spark uses memory as its temporary storage. Read and write process will be faster when using memory than using disks, Spark will be faster than Hadoop.

The Hierarchical Agglomerative Clustering algorithm is implemented in the software. Experiment were done by comparing the execution time of the Hierarchical Agglomerative Clustering algorithm when implemented on Hadoop and Spark. The execution time is recorded for 1GB, 2GB, 3GB, 5GB, 10GB, 15GB, dan 20GB of data.

Based on the experiment, Spark has a faster execution time compared to Hadoop on a large number of partitions. Spark execution time decreases when the number of partitions is increased, whereas Hadoop execution time decreases when the number of partitions is increased. Spark best execution time is still much better than Hadoop best execution time.

**Keywords:** Big Data, Data Reduction, Hierarchical Agglomerative Clustering, Spark, Hadoop





# DAFTAR ISI

<b>DAFTAR ISI</b>	<b>ix</b>
<b>DAFTAR GAMBAR</b>	<b>xi</b>
<b>DAFTAR TABEL</b>	<b>xv</b>
<b>1 PENDAHULUAN</b>	<b>1</b>
1.1 Latar Belakang . . . . .	1
1.2 Rumusan Masalah . . . . .	2
1.3 Tujuan . . . . .	2
1.4 Batasan Masalah . . . . .	2
1.5 Metodologi . . . . .	3
1.6 Sistematika Pembahasan . . . . .	3
<b>2 LANDASAN TEORI</b>	<b>5</b>
2.1 <i>Big Data</i> . . . . .	5
2.2 Algoritma Hierarchical Clustering . . . . .	6
2.3 Hadoop . . . . .	12
2.3.1 Hadoop Distributed File System (HDFS) . . . . .	13
2.3.2 MapReduce . . . . .	15
2.3.3 YARN . . . . .	17
2.4 Spark . . . . .	17
2.4.1 Komponen Spark . . . . .	18
2.4.2 Tiga Cara Membangun Spark di Atas Hadoop . . . . .	20
2.4.3 Arsitektur Spark . . . . .	20
2.4.4 <i>Resilient Distributed Datasets</i> (RDD) . . . . .	21
2.5 Scala . . . . .	24
2.5.1 <i>Expressions</i> . . . . .	24
2.5.2 <i>Blocks</i> . . . . .	25
2.5.3 <i>Loop dan Conditional</i> . . . . .	25
2.5.4 Functions . . . . .	26
2.5.5 Methods . . . . .	27
2.5.6 Class dan Object . . . . .	27
2.5.7 Higher Order Function . . . . .	28
<b>3 STUDI DAN EKSPLORASI APACHE SPARK</b>	<b>31</b>
3.1 Instalasi Apache Spark . . . . .	31
3.2 Eksplorasi Spark Shell . . . . .	33
3.3 Instalasi Apache Spark pada <i>Multi-Node Cluster</i> . . . . .	34
3.4 Percobaan Spark Submit . . . . .	36
<b>4 ANALISIS DAN PERANCANGAN</b>	<b>43</b>
4.1 Analisis Masalah . . . . .	43

4.1.1	Identifikasi Masalah . . . . .	43
4.1.2	Analisis <i>Hierarchical Agglomerative Clustering</i> MapReduce . . . . .	44
4.1.3	Analisis Masukan dan Keluaran . . . . .	46
4.1.4	Diagram Alur . . . . .	48
4.1.5	Analisis <i>Hierarchical Agglomerative Clustering</i> pada Spark . . . . .	51
4.2	Perancangan Perangkat Lunak . . . . .	59
4.2.1	Diagram <i>Use Case</i> dan Skenario . . . . .	59
4.2.2	Diagram Kelas . . . . .	61
4.2.3	Rancangan Antarmuka . . . . .	65
<b>5</b>	<b>IMPLEMENTASI DAN PENGUJIAN PERANGKAT LUNAK</b>	<b>69</b>
5.1	Implementasi Perangkat Lunak . . . . .	69
5.1.1	Lingkungan Perangkat Kerat . . . . .	69
5.1.2	Lingkungan Perangkat Lunak . . . . .	69
5.1.3	User Interface . . . . .	70
5.2	Pengujian Fungsional Perangkat Lunak . . . . .	72
5.3	Hasil Eksperimen Perangkat Lunak . . . . .	74
5.4	Percobaan Dampak Partisi pada Performa Perangkat Lunak Spark dan Hadoop . . . . .	75
<b>6</b>	<b>KESIMPULAN DAN SARAN</b>	<b>99</b>
6.1	Kesimpulan . . . . .	99
6.2	Saran . . . . .	99
	<b>DAFTAR REFERENSI</b>	<b>101</b>
	<b>A KODE PROGRAM</b>	<b>103</b>
	<b>B KODE PROGRAM UNTUK ANTARMUKA</b>	<b>109</b>

## DAFTAR GAMBAR

2.1	Karakteristik <i>big data</i>	5
2.2	Matriks jarak	6
2.3	Matriks jarak	7
2.4	<i>dendrogram</i>	7
2.5	Metode <i>single linkage</i>	8
2.6	Metode <i>complete linkage</i>	8
2.7	Metode <i>centroid linkage</i>	9
2.8	Matriks jarak	9
2.9	Hasil penggabungan <i>cluster</i>	10
2.10	Hasil rekalkulasi	11
2.11	Hasil akhir <i>dendrogram</i>	11
2.12	Perpotongan <i>dendrogram</i>	11
2.13	Modul-modul Hadoop	13
2.14	Arsitektur HDFS	14
2.15	Arsitektur MapReduce	15
2.16	Proses MapReduce	16
2.17	Proses menjalankan aplikasi pada YARN	17
2.18	Komponen pada Spark	18
2.19	Macam-macam cara instalasi Spark	20
2.20	Arsitektur Spark	20
3.1	<i>Spark Shell</i>	32
3.2	<i>Word Count</i>	33
3.3	IntelliJ IDEA	37
3.4	Proyek sbt	38
3.5	Konfigurasi proyek	39
3.6	Struktur proyek	39
3.7	Konfigurasi sbt	40
3.8	<i>object WordCount</i>	40
3.9	Kode WordCount	40
3.10	JAR	41
3.11	Hasil perintah 'sbt package'	41
3.12	Penggumpulan JAR kepada <i>spark-submit</i>	41
3.13	Alamat Spark UI	42
3.14	Spark UI	42
4.1	Penulisan kepada disk di MapReduce	44
4.2	Penulisan kepada memori di Spark	44
4.3	Diagram alur perangkat lunak	48
4.4	Partisi RDD	49
4.5	RDD <i>parsing</i> dan kelas <i>Node</i>	49
4.6	<i>Worker</i> memproses partisi	50
4.7	Pengelompokkan <i>Node</i> berdasarkan <i>key</i>	50

4.8	Proses reduksi dan kelas <i>Pattern</i>	51
4.9	Penyimpanan pola pada HDFS	51
4.10	Contoh perhitungan matriks dan pembentukan dendrogram	58
4.11	Contoh pemotongan <i>dendrogram</i>	59
4.12	Diagram <i>use case</i> perangkat lunak <i>Hierarchical Agglomerative Clustering</i>	60
4.13	Diagram kelas	61
4.14	Kelas Main, SparkConfig, SparkContext	61
4.15	Kelas DataReducer	62
4.16	Kelas Dendrogram	62
4.17	Kelas Cluster	63
4.18	Kelas Pattern	64
4.19	Kelas Node	64
4.20	Rancangan antarmuka menu Jalankan Program	65
4.21	Halaman web Hadoop	66
4.22	Rancangan antarmuka menu Lihat Pola	67
4.23	Rancangan antarmuka halaman partisi	67
4.24	Rancangan antarmuka halaman pola	68
4.25	Halaman web HDFS	68
5.1	Tampilan menu <i>Submit</i>	70
5.2	Tampilan menu <i>Data</i>	70
5.3	Tampilan halaman sesudah <i>submit</i>	71
5.4	Tampilan halaman <i>list</i>	71
5.5	Tampilan halaman data	72
5.6	Hasil Percobaan Partisi Spark dan Hadoop dengan ukuran data 1GB, jumlah objek maksimum 30, dan total 10 core	76
5.7	Hasil Percobaan Partisi Spark dan Hadoop dengan ukuran data 2GB, jumlah objek maksimum 30, dan total 10 core	77
5.8	Hasil Percobaan Partisi Spark dan Hadoop dengan ukuran data 3GB, Objek Maksimum 30, dan Total 10 Core	78
5.9	Hasil Percobaan Partisi Spark dan Hadoop dengan ukuran data 5GB, Objek Maksimum 30, dan Total 10 Core	79
5.10	Hasil Percobaan Partisi Spark dan Hadoop dengan ukuran data 10GB, Objek Maksimum 30, dan Total 10 Core	81
5.11	Hasil Percobaan Partisi Spark dan Hadoop dengan ukuran data 15GB, Objek Maksimum 30, dan Total 10 Core	83
5.12	Hasil Percobaan Partisi Spark dan Hadoop dengan ukuran data 20GB, Objek Maksimum 30, dan Total 10 Core	84
5.13	Hasil Percobaan Jumlah Partisi Spark dan Hadoop dengan ukuran data 5GB, Objek Maksimum 50, dan Total 10 Core	86
5.14	Hasil Percobaan Jumlah Partisi Spark dan Hadoop dengan Ukuran Data 10 GB, Objek Maksimum 50, dan Total 10 Core	87
5.15	Hasil Percobaan Jumlah Partisi Spark dan Hadoop dengan Ukuran Data 15 GB, Objek Maksimum 50, dan Total 10 Core	88
5.16	Hasil Percobaan Jumlah Partisi Spark dan Hadoop dengan Ukuran Data 20 GB, Objek Maksimum 50, dan Total 10 Core	89
5.17	Hasil Percobaan Jumlah Partisi Spark dan Hadoop dengan Ukuran Data 5 GB, Objek Maksimum 100, dan Total 10 Core	90
5.18	Hasil Percobaan Jumlah Partisi Spark dan Hadoop dengan Ukuran Data 10 GB, Objek Maksimum 100, dan Total 10 Core	91
5.19	Hasil Percobaan Jumlah Partisi Spark dan Hadoop dengan Ukuran Data 15 GB, Objek Maksimum 100, dan Total 10 Core	93

5.20 Hasil Percobaan Jumlah Partisi Spark dan Hadoop dengan Ukuran Data 20 GB, Objek Maksimum 100, dan Total 10 Core . . . . .	94
5.21 Hasil Percobaan Jumlah Partisi Spark dan Hadoop dengan Ukuran Data 10 GB, Objek Maksimum 30, dan Total 10 Core . . . . .	95
5.22 Hasil Percobaan Jumlah Partisi Spark dan Hadoop dengan Ukuran Data 10 GB, Objek Maksimum 30, dan Total 10 Core . . . . .	96



## DAFTAR TABEL

2.1	Tabel Data Koordinat . . . . .	9
2.2	Tabel Contoh Data <i>Cluster</i> . . . . .	12
2.3	Tabel Hasil Pola Cluster A . . . . .	12
2.4	Tabel transformations . . . . .	23
2.5	Tabel Actions . . . . .	24
5.1	Tabel data yang digunakan pada eksperimen . . . . .	75
5.2	Percobaan Jumlah Partisi Hadoop dan Spark dengan Ukuran Data 1 GB . . . . .	75
5.3	Percobaan Jumlah Partisi Hadoop dan Spark dengan Ukuran Data 2 GB . . . . .	76
5.4	Percobaan Jumlah Partisi Hadoop dan Spark dengan Ukuran Data 3 GB . . . . .	78
5.5	Percobaan Jumlah Partisi Hadoop dan Spark dengan Ukuran Data 5 GB . . . . .	79
5.6	Percobaan Jumlah Partisi Spark dengan Ukuran Data 10 GB . . . . .	80
5.7	Percobaan Jumlah Partisi Hadoop dengan Ukuran Data 10 GB . . . . .	80
5.8	Percobaan Jumlah Partisi Spark dengan Ukuran Data 15 GB . . . . .	82
5.9	Percobaan Jumlah Partisi Hadoop dengan Ukuran Data 15 GB . . . . .	82
5.10	Percobaan Jumlah Partisi Spark dengan Ukuran Data 20 GB . . . . .	84
5.11	Percobaan Jumlah Partisi Hadoop dengan Ukuran Data 20 GB . . . . .	84
5.12	Percobaan Jumlah Partisi Hadoop dengan Ukuran Data 5 GB . . . . .	85
5.13	Percobaan Jumlah Partisi Spark dengan Ukuran Data 5 GB . . . . .	85
5.14	Percobaan Jumlah Partisi Hadoop dengan Ukuran Data 10 GB . . . . .	86
5.15	Percobaan Jumlah Partisi Spark dengan Ukuran Data 10 GB . . . . .	87
5.16	Percobaan Jumlah Partisi Hadoop dengan Ukuran Data 15 GB . . . . .	88
5.17	Percobaan Jumlah Partisi Spark dengan Ukuran Data 15 GB . . . . .	88
5.18	Percobaan Jumlah Partisi Hadoop dengan Ukuran Data 20 GB . . . . .	89
5.19	Percobaan Jumlah Partisi Spark dengan Ukuran Data 20 GB . . . . .	89
5.20	Percobaan Jumlah Partisi Hadoop dengan Ukuran Data 5 GB . . . . .	90
5.21	Percobaan Jumlah Partisi Spark dengan Ukuran Data 5 GB . . . . .	90
5.22	Percobaan Jumlah Partisi Hadoop dengan Ukuran Data 10 GB . . . . .	91
5.23	Percobaan Jumlah Partisi Spark dengan Ukuran Data 10 GB . . . . .	91
5.24	Percobaan Jumlah Partisi Hadoop dengan Ukuran Data 15 GB . . . . .	92
5.25	Percobaan Jumlah Partisi Spark dengan Ukuran Data 15 GB . . . . .	92
5.26	Percobaan Jumlah Partisi Hadoop dengan Ukuran Data 20 GB . . . . .	93
5.27	Percobaan Jumlah Partisi Spark dengan Ukuran Data 20 GB . . . . .	94
5.28	Percobaan Jumlah Partisi Hadoop dengan Ukuran Data 10 GB . . . . .	95
5.29	Percobaan Jumlah Partisi Spark dengan Ukuran Data 10 GB . . . . .	95
5.30	Percobaan Jumlah Partisi Hadoop dengan Ukuran Data 10 GB . . . . .	96
5.31	Percobaan Jumlah Partisi Spark dengan Ukuran Data 10 GB . . . . .	96





# BAB 1

## PENDAHULUAN

### 1.1 Latar Belakang

*Big data* adalah sebuah istilah yang menggambarkan volume data yang besar, baik data yang terstruktur maupun data yang tidak terstruktur. Data-data tersebut memiliki potensi untuk digali menjadi informasi yang penting. Dalam bidang *big data* ada beberapa tantangan seperti volume data yang besar, kecepatan aliran data yang masuk, dan variasi data dengan format yang berbeda. Tantangan tersebut membuat aplikasi pemrosesan data tradisional tidak bisa memproses dan menganalisis *big data*. Muncul teknologi-teknologi seperti Hadoop dan Spark yang dirancang khusus untuk menangani *big data*.

*Big data* akan lebih mudah dianalisis dan diterapkan teknik-teknik *data-mining* ketika volume *big data* tersebut telah direduksi. Dengan mereduksi data, kita bisa menghemat biaya pengiriman data, *disk space*, dan jumlah data yang diproses. Hasil dari reduksi *big data* harus bisa mewakili data yang belum direduksi secara akurat.

Salah satu cara mereduksi data adalah dengan menggunakan algoritma *Hierarchical Agglomerative Clustering*. Algoritma tersebut cocok untuk data yang tidak memiliki atribut yang terlalu banyak. Journal ilmiah berjudul *Big Data Reduction Technique using Parallel Hierarchical Agglomerative Clustering* menjabarkan algoritma *Hierarchical Agglomerative Clustering* berbasis MapReduce pada Hadoop [1]. Penelitian tersebut membuktikan bahwa data yang direduksi dengan algoritma tersebut bisa mewakili data yang belum direduksi. Algoritma *Hierarchical Agglomerative Clustering* bekerja dengan mengubah setiap objek menjadi *sub-cluster*. Kemudian, *sub-cluster* akan digabung dengan *sub-cluster* lainnya secara bertahap berdasarkan jarak antara *sub-cluster* sampai terbentuknya sebuah *cluster*. *Cluster* tersebut akan menjadi akar dari hierarki.

Meskipun hasil reduksi data dengan algoritma *Hierarchical Agglomerative Clustering* berbasis MapReduce pada Hadoop dapat mewakili data yang belum direduksi secara akurat, MapReduce pada Hadoop memiliki kekurangan. Hadoop tidak efisien dalam melakukan proses iterasi, *intermediate data* tidak dapat disimpan pada memori. Hadoop perlu melakukan penulisan dan pembacaan kepada *disk* di antara setiap tahap Map dan Reduce.

Spark adalah *distributed cluster-computing framework* yang bisa menggantikan MapReduce beserta kekurangannya. *In-memory processing* pada Spark dapat mengalahkan kecepatan pemrosesan pada Hadoop MapReduce. Karena data disimpan pada RAM, kecepatan pemrosesan akan jauh lebih cepat. Spark membaca data yang akan direduksi dari RAM. Pembacaan data dari RAM akan lebih cepat dibanding disk.

Pada skripsi ini, dibangun sebuah perangkat lunak yang dapat mereduksi *big data*. Perangkat lunak tersebut akan dibangun menggunakan *framework* terdistribusi Spark dan mengimplementasikan algoritma *Hierarchical Agglomerative Clustering* yang khusus dirancang untuk lingkungan Spark. Perangkat lunak dapat menampilkan hasil reduksi dalam format tabel. Dengan menggunakan Spark, waktu proses reduksi

- 1 data menjadi lebih cepat dibanding MapReduce.

## 2 **1.2 Rumusan Masalah**

3 Berdasarkan latar belakang di atas, dapat dibentuk rumusan masalah sebagai berikut:

- 4 1. Bagaimana cara kerja algoritma *Hierarchical Agglomerative Clustering* berbasis MapReduce untuk  
5 mereduksi *big data*?
- 6 2. Bagaimana cara mengkustomisasi dan mengimplementasikan algoritma *Agglomerative Clustering*  
7 pada sistem tersebar Spark?
- 8 3. Bagaimana mengukur kinerja hasil dari implementasi dari algoritma *Agglomerative Clustering* pada  
9 sistem tersebar Spark?
- 10 4. Bagaimana cara mempresentasikan data yang telah direduksi?

## 11 **1.3 Tujuan**

12 Berdasarkan rumusan masalah di atas, tujuan dari penelitian adalah sebagai berikut:

- 13 1. Mempelajari cara kerja algoritma *Hierarchical Agglomerative Clustering* berbasis MapReduce untuk  
14 mereduksi *big data*.
- 15 2. Mengkustomisasi dan mengimplementasikan algoritma *Hierarchical Agglomerative Clustering* pada  
16 lingkungan Spark.
- 17 3. Melakukan eksperimen pada lingkungan sistem tersebar Spark untuk mengukur kinerja algoritma  
18 lingkungan Spark.
- 19 4. Membuat modul program untuk menginterpretasikan data yang telah direduksi.

## 20 **1.4 Batasan Masalah**

21 Batasan masalah pada skripsi ini adalah sebagai berikut:

- 22 1. Studi literatur Hadoop hanya dilakukan pada dasar dan file system Hadoop yaitu HDFS.
- 23 2. Studi literatur Apache Spark hanya mempelajari konsep dasar dari Apache Spark, *Resilient Distributed*  
24 *Dataset* (RDD), dan implementasi algoritma *Hierarchical Agglomerative Clustering* (HAC).
- 25 3. Metode reduksi data yang dibahas secara mendalam hanya metode *agglomerative clustering*.
- 26 4. Algoritma *Hierarchical Agglomerative Clustering* diimplementasikan secara paralel pada sistem  
27 terdistribusi Spark.

## 1.5 Metodologi

Metodologi yang digunakan dalam pembuatan skripsi ini adalah:

1. Melakukan studi literatur Hadoop hanya mempelajari konsep dasar dari Hadoop dan *Hadoop Distributed File System* (HDFS).
2. Melakukan studi literatur tentang konsep Apache Spark.
3. Melakukan studi literatur bahasa pemrograman Scala.
4. Melakukan studi literatur tentang algoritma *Hierarchical Agglomerative Clustering*.
5. Melakukan instalasi dan konfigurasi Apache Spark.
6. Melakukan eksperimen dengan bahasa pemrograman Scala.
7. Melakukan eksperimen dengan Spark RDD.
8. Melakukan kustomisasi algoritma *Hierarchical Agglomerative Clustering* untuk Spark.
9. Mencari dan mengumpulkan data uji coba yang bervolume besar.
10. Merancang dan mengimplementasikan perangkat lunak.
11. Melakukan eksperimen terhadap perangkat lunak dan menganalisis hasil eksperimen.
12. Menulis dokumen skripsi.

## 1.6 Sistematika Pembahasan

Laporan penelitian tersusun ke dalam enam bab secara sistematis sebagai berikut:

- Bab 1 Pendahuluan  
Berisi latar belakang, rumusan masalah, tujuan, batasan masalah, metodologi penelitian, dan sistematika pembahasan.
- Bab 2 Dasar Teori  
Berisi dasar teori tentang *big data*, *Hierarchical Agglomerative Clustering*, Hadoop, Spark, dan Scala.
- Bab 3 Studi dan Eksplorasi Apache Spark  
Berisi percobaan-percobaan yang dilakukan pada Spark.
- Bab 4 Analisis dan Perancangan  
Berisi analisis masalah, diagram alur, *use case* dan skenario, diagram kelas, dan perancangan antarmuka.
- Bab 5 Implementasi dan Pengujian  
Berisi implementasi antarmuka perangkat lunak, pengujian eksperimen, dan kesimpulan dari pengujian.
- Bab 5 Implementasi dan Pengujian  
Berisi kesimpulan awal sampai akhir penelitian dan saran untuk penelitian selanjutnya.

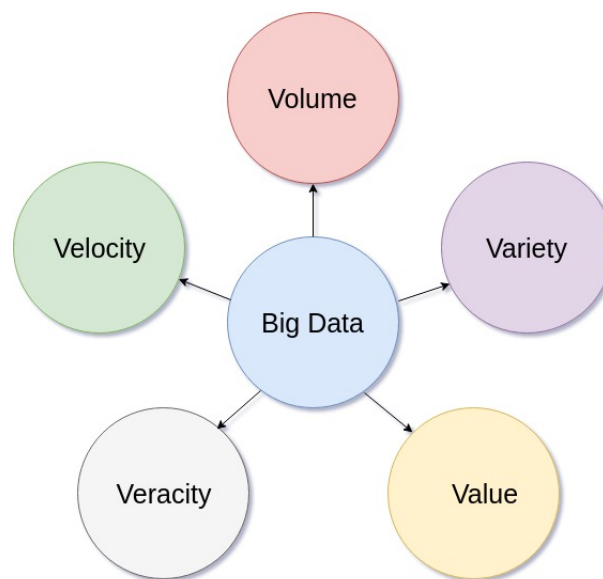


## BAB 2

### LANDASAN TEORI

#### 2.1 *Big Data*

*Big data* adalah istilah yang menggambarkan kumpulan data dalam jumlah yang sangat besar, baik data yang terstruktur maupun data yang tidak terstruktur. Kumpulan data tersebut menyimpan informasi yang bisa dianalisis dan diproses untuk memberikan wawasan kepada organisasi atau perusahaan. Data-data tersebut berasal dari satu atau lebih sumber dengan kecepatan yang tinggi dan *format* yang berbeda-beda. Karena ukurannya dan keberagaman data, *big data* menjadi sulit untuk ditangani atau diproses jika hanya menggunakan manajemen basis data atau aplikasi pemrosesan data tradisional [2].



Gambar 2.1: Karakteristik *big data*

Berdasarkan Gambar 2.1, *big data* memiliki lima karakteristik di antaranya [2]:

1. *Volume*: *big data* memiliki jumlah data yang sangat besar sehingga dalam proses pengolahan data dibutuhkan suatu penyimpanan yang besar dan dibutuhkan analisis yang lebih spesifik.
2. *Velocity*: *big data* memiliki aliran data yang sangat cepat. Data baru dihasilkan dengan kecepatan yang tinggi dari satu atau lebih sumber.
3. *Variety*: *big data* memiliki bentuk format data yang beragam, baik terstruktur ataupun tidak terstruktur dan bergantung pada banyaknya sumber data. Data dapat berupa gambar, video dan tipe data lainnya.

4. *Veracity*: *big data* dapat mengandung data yang tidak akurat atau rusak. Kualitas data dalam *big data* bisa berbeda-beda bergantung pada sumber. Analisis *big data* akan sangat dipengaruhi dengan keakuratan data.

5. *Value*: *big data* harus memiliki *value*. Tidak ada gunanya bila kita memiliki akses terhadap *big data*, tetapi data-data tersebut tidak memiliki nilai apapun. Data yang tidak memiliki nilai adalah data yang tidak berguna dan memakan biaya untuk disimpan.

*Big data* sangat bermanfaat ketika diterapkan di berbagai macam bidang seperti bisnis, kesehatan, pemerintahan, pertanian dan lainnya. Ketika organisasi mampu menggabungkan jumlah data besar yang dimilikinya dengan analisis bertenaga tinggi, organisasi dapat menyelesaikan tantangan dan masalah yang berhubungan dengan bisnis seperti:

1. Menentukan akar penyebab kegagalan untuk setiap masalah bisnis.
2. Menghasilkan informasi mengenai titik penting penjualan berdasarkan kebiasaan pelanggan dalam membeli.
3. Menghitung kembali seluruh risiko yang ada dalam waktu yang singkat.
4. Mendeteksi perilaku penipuan yang dapat mempengaruhi organisasi.

## 2.2 Algoritma Hierarchical Clustering

*Hierarchical Clustering Algorithm* (HCA) adalah metode analisis kelompok yang berusaha untuk membangun sebuah hierarki dengan mengelompokkan data. Dengan mengelompokkan data-data tersebut, data pada kelompok yang sama memiliki kemiripan yang tinggi dan data pada kelompok yang berbeda memiliki kemiripan yang rendah [1]. Dalam reduksi data, *cluster* yang merepresentasikan data-data pada *cluster* tersebut akan digunakan untuk mengganti data-data mentah [1]. Seberapa efektif cara ini bergantung pada sifat data yang ditangani. Data-data yang bisa dikelompokkan ke dalam *cluster* yang berbeda akan sangat cocok dengan cara ini [1]. Pada dasarnya HCA dibagi menjadi dua jenis, yaitu *agglomerative (bottom-up)* dan *devisive (top-down)* [1]. Pendekatan *agglomerative* berusaha membentuk sebuah hierarki dengan menggabungkan *cluster*. Setiap objek akan dimasukkan kepada *cluster* tersendiri. Sebaliknya, pendekatan *devisive* akan berusaha memecah *cluster* untuk membentuk sebuah hierarki. Setiap objek berada pada satu *cluster* pada awalnya dan akan dipecah kepada *cluster* yang berbeda.

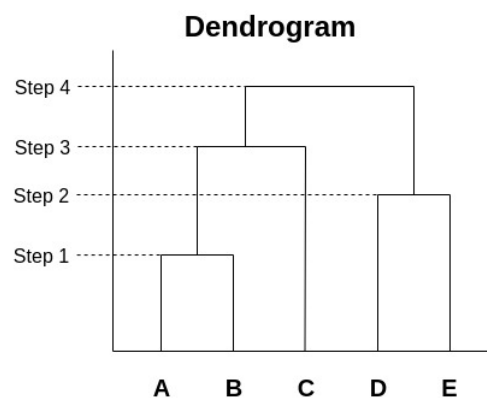
	A	B	C	D
A	0	2	5	6
B	2	0	7	3
C	5	7	0	4
D	6	3	4	0

Gambar 2.2: Matriks jarak

	(A,B)	C	D
(A,B)	0	?	?
C	?	0	4
D	?	4	0

Gambar 2.3: Matriks jarak

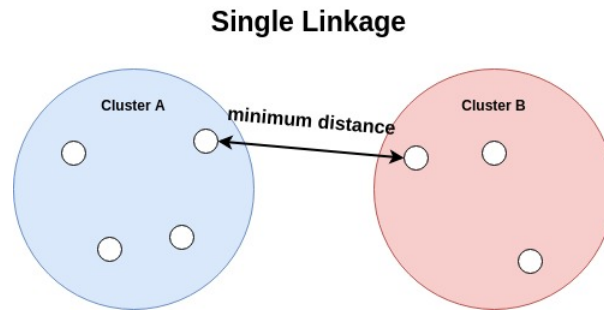
Pada *Hierarchical Agglomerative Clustering*, awalnya setiap objek akan dimasukkan kepada *cluster* tersendiri. Matriks jarak digunakan untuk merepresentasikan jarak antara *cluster*. Kemudian, dua buah *cluster* yang memiliki jarak terdekat akan digabungkan menjadi satu *cluster*. Jarak antara *cluster* dapat dihitung dengan tiga metode, yaitu *single linkage*, *complete linkage*, dan *centroid linkage* [3]. Pada Gambar 2.2, *cluster* A dan *cluster* B akan digabung menjadi satu karena jarak antara keduanya adalah terkecil dibanding dengan yang lainnya. Gambar 2.3 adalah hasil dari penggabungan *cluster* A dan *cluster* B. Kemudian, matriks jarak perlu dihitung kembali untuk mencari jarak baru antara *cluster* baru dengan *cluster* lainnya. Penggabungan *cluster* akan diulangi sampai tersisa satu *cluster*. *Hierarchical Agglomerative Clustering* akan hanya membutuhkan maksimal  $n$  iterasi. Hasil dari penggabungan *cluster* adalah sebuah hierarki. *Dendrogram* sangat umum digunakan untuk menggambarkan proses *Hierarchical Agglomerative Clustering*. Contoh *dendrogram* dapat dilihat pada Gambar 2.4.



Gambar 2.4: dendrogram

Berikut adalah penjelasan mengenai metode *single linkage*, *complete linkage*, dan *centroid linkage*:

- *Single linkage*: metode ini mencari jarak minimum dari perbandingan setiap anggota antara dua buah *cluster*. Bila terdapat *cluster* A dan *cluster* B, maka setiap anggota pada *cluster* A akan dihitung jaraknya kepada setiap anggota pada *cluster* B. Kemudian jarak minimum antara anggota akan diambil sebagai hasilnya. Untuk menghitung jarak antara anggota dapat digunakan *euclidean distance*, *manhattan distance*, atau ruang metrik lainnya. Ruang metrik yang digunakan disesuaikan dengan kebutuhan dan atribut dari data. Contoh *single linkage* dapat dilihat pada Gambar 2.5.

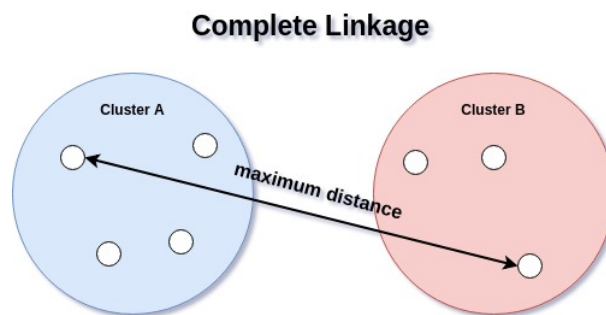
Gambar 2.5: Metode *single linkage*

Rumus 2.1 adalah rumus untuk *single linkage*:

$$\min\{d(a,b) : a \in A, b \in B\}, \quad (2.1)$$

dengan  $a$  dan  $b$  merupakan anggota dari *cluster* A dan B.

- *Complete linkage*: metode ini adalah kebalikan dari metode *single linkage*. Bila terdapat *cluster* A dan B, maka setiap anggota pada *cluster* A akan dihitung jaraknya kepada setiap anggota pada *cluster* B. Kemudian jarak maksimum antara anggota akan diambil sebagai hasilnya. Contoh *complete linkage* dapat dilihat pada Gambar 2.6.

Gambar 2.6: Metode *complete linkage*

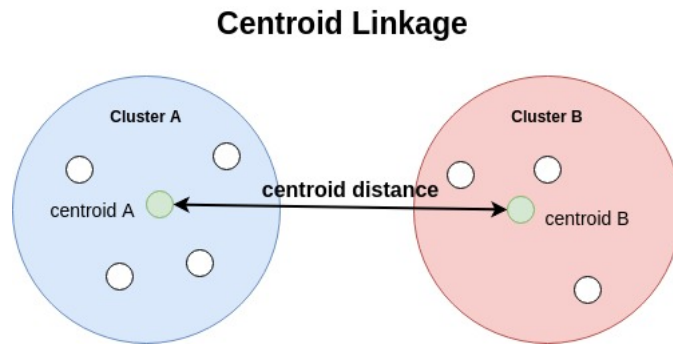
Rumus 2.2 adalah rumus untuk *complete linkage*:

$$\max\{d(a,b) : a \in A, b \in B\}, \quad (2.2)$$

dengan  $a$  dan  $b$  merupakan anggota dari *cluster* A dan B.

- *Centroid linkage*: metode ini menghitung jarak antara *centroid* dari dua buah *cluster*. *Centroid* merupakan titik tengah dari sebuah *cluster*. *Centroid* sebuah *cluster* didapatkan dengan menghitung rata-rata dari setiap atribut dari anggota pada *cluster*. Contoh *centroid linkage* dapat dilihat pada Gambar 2.7.



Gambar 2.7: Metode *centroid linkage*

Rumus 2.3 adalah rumus untuk *centroid linkage*:

$$\|c_a - c_b\|, \quad (2.3)$$

dengan  $c_a$  dan  $c_b$  merupakan *centroid* dari *cluster A* dan *B*.

Tabel 2.1: Tabel Data Koordinat

Cluster	x	y
A	2	2
B	2	3
C	4	6
D	8	10

Sebagai contoh, diberikan data yang memiliki atribut berupa koordinat  $x$  dan  $y$ . Data dapat dilihat pada Tabel 2.1. Data tersebut akan diolah dengan algoritma *Hierarchical Agglomerative Clustering* menggunakan metode *single linkage* dan *euclidean distance* untuk menghitung jaraknya antara anggotanya. Berikut adalah langkah-langkah penyelesaiannya.

- Pertama, hitung matriks jarak antara *cluster*. Karena setiap *cluster* hanya memiliki satu anggota pada awalnya, Jarak antara *cluster* dapat langsung dihitung menggunakan *euclidean distance*. Matriks jarak yang dihasilkan bisa dilihat pada Gambar 2.8.

	A	B	C	D
A	0	1.0	4.47	10.0
B	1.00	0	3.61	9.22
C	4.47	3.61	0	5.66
D	10.0	9.22	5.66	0

Gambar 2.8: Matriks jarak

Jarak antara *cluster* A dan *cluster* B dapat dihitung dengan cara berikut:

$$\begin{aligned}
 d &= \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \\
 &= \sqrt{(2 - 2)^2 + (2 - 3)^2} \\
 &= \sqrt{0 + 1} \\
 &= \sqrt{1} \\
 &= 1
 \end{aligned} \tag{2.4}$$

2. Selanjutnya, gabungkan dua *cluster* yang memiliki jarak terdekat. Pada contoh ini, *cluster* A yang dibandingkan terhadap *cluster* B memiliki nilai terkecil yaitu 1. Jarak antara kedua *cluster* adalah yang terdekat. Hasil dari penggabungan kedua *cluster* dapat dilihat pada Gambar 2.9.

	(A, B)	C	D
(A, B)	0	?	?
C	?	0	5.66
D	?	5.66	0

Gambar 2.9: Hasil penggabungan *cluster*

3. Setelah itu, matriks jarak harus dihitung ulang untuk mencari jarak antara *cluster* barunya, yaitu (A,B) dengan yang lainnya. Untuk menghitung ulang antara *cluster* baru dengan *cluster* lainnya, digunakan metode *single linkage*. Pada tahap ini setiap anggota dari *cluster* (A,B) akan dihitung jaraknya terhadap *cluster* C dan *cluster* D. Nilai minimum akan diambil sebagai hasil perbandingannya karena metode yang digunakan adalah *single linkage*. Berikut adalah contoh perhitungan antara *cluster* (A,B) dengan *cluster* C menggunakan metode *single linkage*.

$$\begin{aligned}
 d(A,C) &= \sqrt{(2 - 2)^2 + (4 - 6)^2} \\
 &= 4.47
 \end{aligned} \tag{2.5}$$

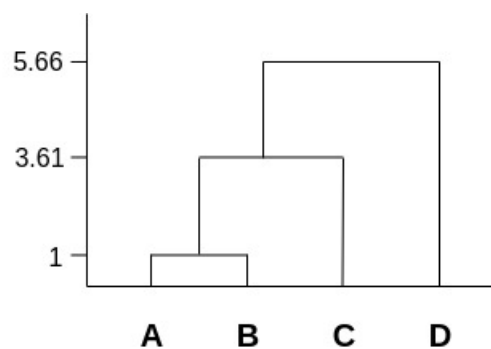
$$\begin{aligned}
 d(B,C) &= \sqrt{(2 - 3)^2 + (4 - 6)^2} \\
 &= 3.61
 \end{aligned} \tag{2.6}$$

Berdasarkan perhitungan di atas, nilai 3.61 diambil sebagai hasil karena nilai tersebut lebih kecil dibandingkan 4.47. Contoh hasil dapat dilihat pada Gambar 2.10.

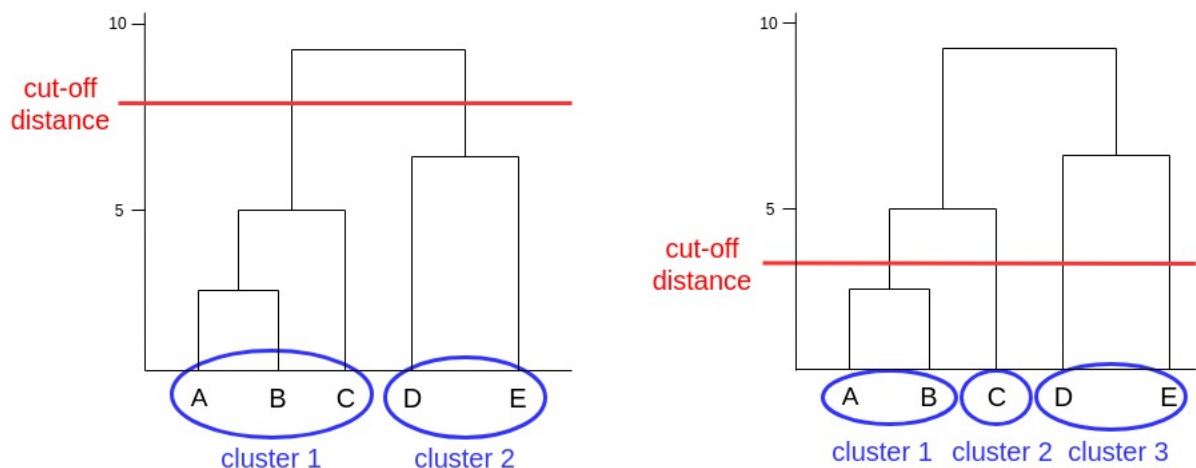
	(A, B)	C	D
(A, B)	0	3.61	?
C	3.61	0	5.66
D	?	5.66	0

Gambar 2.10: Hasil rekalkulasi

- 1 4. Ulangi langkah 2 dan 3 sampai satu *cluster* yang tersisa. Hasil akhir dalam bentuk *dendrogram* dapat  
 2 dilihat pada Gambar 2.11.

Gambar 2.11: Hasil akhir *dendrogram*

- 3 Setelah *dendrogram* terbentuk, *dendrogram* perlu dipotong berdasarkan nilai *cut-off distance* yang diten-  
 4 tukan. Nilai *cut-off distance* menentukan banyaknya *cluster* yang dihasilkan ketika memotong *dendrogram*.  
 5 Semakin tinggi nilai *cut-off distance*, semakin sedikit *cluster* yang dihasilkan dan sebaliknya. Berdasarkan  
 6 Gambar 2.11, dapat dilihat bahwa nilai *cut-off distance* yang lebih tinggi menghasilkan *cluster-cluster* yang  
 7 lebih sedikit. Sedangkan, nilai *cut-off distance* yang lebih rendah menghasilkan *cluster-cluster* yang lebih  
 8 banyak. Perpotongan akan berdampak kepada hasil akhir ukuran data yang dihasilkan.

Gambar 2.12: Perpotongan *dendrogram*

- 9 Dari setiap *cluster* yang dihasilkan dari perpotongan, perlu dicari jumlah anggota pada *cluster*, nilai

1 minimum, maksimum, rata-rata, dan standar deviasi dari setiap atribut. Nilai-nilai tersebut dapat disebut  
 2 sebagai pola. Pola ini akan merepresentasikan dan menggambarkan karakteristik *cluster* tersebut. Pola ini  
 3 akan disimpan sebagai hasil akhir untuk menggantikan data aslinya. Sebagai contoh diberikan sebuah *cluster*  
 4 A yang memiliki 4 anggota pada *cluster*-nya. Setiap anggota memiliki 2 nilai atribut yang berbeda. Data  
 5 untuk *cluster* A dapat dilihat pada Tabel 2.2. Data pada tabel ini akan digunakan untuk mencari pola untuk  
 6 *cluster* A.

Tabel 2.2: Tabel Contoh Data *Cluster*

Cluster A		
No	atribut 1	atribut 2
1	2	1
2	4	5
3	5	10
4	6	7

7 Hasil pola dari *cluster* A dapat dilihat pada Tabel 2.3.

Tabel 2.3: Tabel Hasil Pola Cluster A

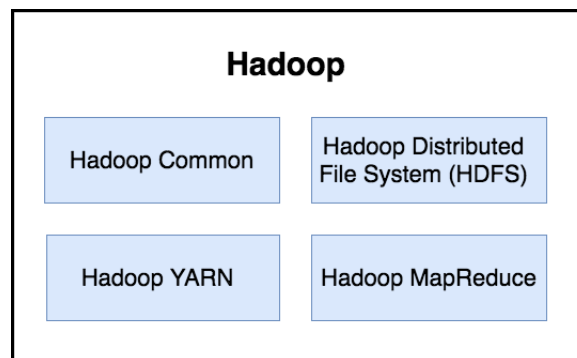
jumlah anggota pada cluster	4	
	atribut 1	atribut 2
nilai minimum	2	1
nilai maksimum	6	10
nilai rata-rata	4.25	5.75
standar deviasi	1.479	3.269

## 8 2.3 Hadoop

9 Hadoop dikembangkan oleh Doug Cutting dan Mike Cafarella pada tahun 2005 yang saat itu bekerja di  
 10 Yahoo. Nama Hadoop diberikan berdasarkan mainan 'Gajah' anak dari Doug Cutting. Hadoop adalah  
 11 sebuah *framework* atau platform *open source* berbasis Java. Hadoop memiliki kemampuan untuk menyimpan  
 12 dan memproses data dengan skala yang besar secara terdistribusi pada *cluster*. *Cluster* tersebut terdiri atas  
 13 perangkat keras komoditas [4]. Hadoop menggunakan teknologi Google MapReduce dan *Google File System*  
 14 (GFS) sebagai fondasinya [5]. Beberapa karakteristik yang dimiliki Hadoop adalah sebagai berikut:

- 15 1. *Open Source*: Hadoop merupakan proyek *open source* dan kodenya bisa dimodifikasi sesuai kebutuhan.
- 16 2. *Distributed Computing*: Data disimpan secara terdistribusi pada *Hadoop Distributed File System*  
 17 (HDFS) dan data dapat diproses secara paralel pada *node-node* di *cluster*.
- 18 3. *Fast*: Hadoop sangat cocok untuk melakukan *batch processing* bervolume besar karena mampu  
 19 melakukannya secara paralel.
- 20 4. *Fault Tolerance*: Hadoop melakukan duplikasi data di beberapa *node* yang berbeda. Ketika sebuah node  
 21 gagal memproses data, *node* yang memiliki duplikat data dapat menggantikannya untuk memproses  
 22 data tersebut.
- 23 5. *Reliability*: Kegagalan mesin bukan masalah bagi Hadoop karena adanya duplikasi data.

6. *High Availability*: Data dapat diambil dari sumber yang lain meskipun kegagalan mesin karena adanya duplikasi data.
7. *Scalability*: Hadoop dapat menambahkan node yang lebih banyak ke dalam *cluster* dengan mudah.
8. *Flexibility*: Hadoop dapat menangani data terstruktur maupun data tidak terstruktur.
9. *Economic And Cost Effective*: Hadoop tidak mahal karena berjalan pada *cluster* yang terdiri atas perangkat keras komoditas.
10. *Easy To Use*: Hadoop mempermudah pengguna dalam merancang program paralel. Hadoop sudah menangani pembagian dan penugasan kerja secara paralel.
11. *Data Locality*: Algoritma MapReduce akan didekatkan kepada *cluster* dan tidak sebaliknya. Ukuran data yang besar lebih sulit untuk dipindahkan dibanding ukuran algoritma yang kecil.



Gambar 2.13: Modul-modul Hadoop

Berdasarkan Gambar 2.13, *framework* Apache Hadoop terdiri dari beberapa modul. Modul-modul tersebut membentuk dan membantu pemrosesan data berskala besar. Modul-modul tersebut di antaranya adalah [5]:

1. *Hadoop Common*: modul ini terdiri atas *library* dan *tools* yang dibutuhkan module Hadoop lainnya.
2. *Hadoop Distributed File System (HDFS)*: sebuah file sistem terdistribusi milik Hadoop untuk penyimpanan data.
3. *Hadoop YARN*: *resource-management platform* yang bertanggung jawab untuk mengatur sumber daya pada *cluster*.
4. MapReduce: sebuah model pemrograman untuk pemrosesan skala besar.

### 2.3.1 Hadoop Distributed File System (HDFS)

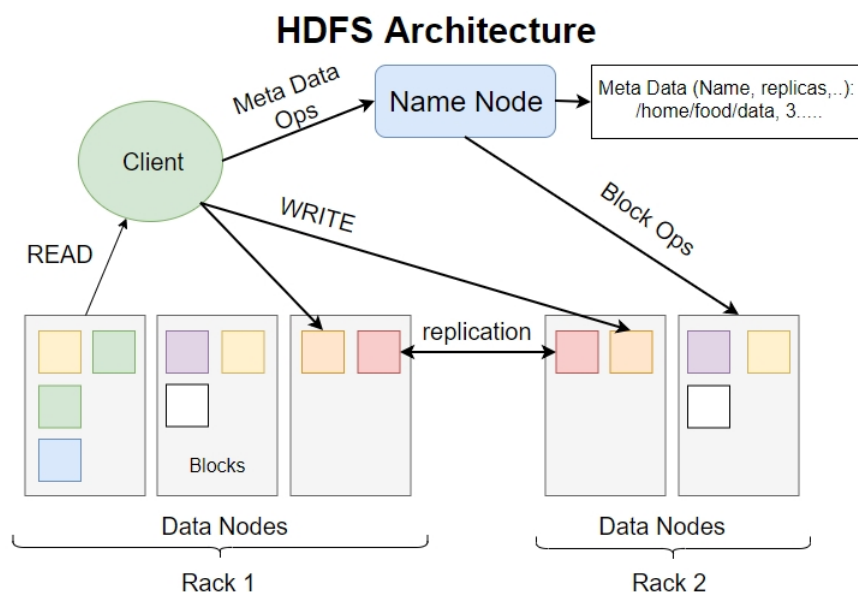
Hadoop Distributed File System (HDFS) adalah sistem file terdistribusi yang dirancang untuk berjalan pada perangkat keras komoditas [5]. HDFS berbeda dari file sistem terdistribusi lainnya karena sifat *fault tolerance* yang tinggi dan dirancang untuk digunakan pada perangkat keras biasa. HDFS menyediakan akses *throughput* yang tinggi dan cocok untuk *data set* yang besar. HDFS awalnya dibangun sebagai infrastruktur untuk proyek

1 mesin pencari web Apache Nutch.

2

3 Kegagalan perangkat keras sudah biasa terjadi. HDFS mungkin terdiri atas ratusan atau ribuan mesin  
 4 server, masing-masing menyimpan bagian data dari file sistem. Faktanya, ada sejumlah besar komponen dan  
 5 setiap komponen memiliki probabilitas kegagalan. Hal ini menandakan bahwa terdapat beberapa komponen  
 6 HDFS selalu tidak berfungsi. Oleh karena itu, deteksi kesalahan dan pemulihan otomatis yang cepat dari  
 7 sistem adalah tujuan arsitektur inti dari HDFS.

8



Gambar 2.14: Arsitektur HDFS

9 Hadoop meimplementasikan arsitektur *Master Slave* pada komponen primernya yaitu HDFS dan Map-  
 10 Reduce [5]. Berdasarkan (Gambar 2.14), *master node* atau disebut NameNode bertugas untuk mengatur  
 11 operasi-operasi seperti membuka, menutup, dan menamakan kembali file atau direktori pada sistem file.  
 12 Selain itu, NameNode meregulasi akses pengguna terhadap file dan mengatur blok mana yang akan diolah  
 13 oleh DataNode [5]. NameNode membuat semua keputusan terkait replikasi blok. NameNode secara berkala  
 14 menerima *heartbeat* dan *block report* dari masing-masing DataNode di *cluster*. *Heartbeat* mengimplikasikan  
 15 bahwa DataNode berfungsi dengan benar.

16

17 *Slave node* atau dapat disebut DataNode merupakan pekerja dari HDFS [5]. DataNode bertanggung  
 18 jawab untuk menjalankan perintah membaca dan menulis untuk file sistem Hadoop. NameNode dapat  
 19 membuat, menghapus, dan mereplikasi blok ketika diberi instruksi dari *master node*. DataNode menyimpan  
 20 dan mengambil blok ketika diperintahkan oleh NameNode. Selain itu, DataNode melaporkan daftar blok-blok  
 21 yang disimpan kepada NameNode secara rutin.

22

23 HDFS dirancang untuk menyimpan file yang berukuran sangat besar di seluruh mesin dalam *cluster* yang  
 24 besar [5]. HDFS menyimpan setiap file sebagai blok yang berurutan. Semua blok dalam file kecuali blok  
 25 terakhir memiliki ukuran yang sama. Bisa dilihat pada Gambar 2.14 bahwa blok-blok file direplikasi untuk

memiliki *fault tolerance* yang tinggi. Ukuran blok dan banyaknya replika dapat dikonfigurasi untuk setiap file. Faktor replikasi dapat ditentukan pada waktu pembuatan file dan dapat diubah nantinya.

Berikut adalah perintah-perintah dasar yang dapat digunakan untuk HDFS [6]:

- Perintah untuk membuat direktori HDFS untuk penyimpanan file.

```
$ hadoop fs -mkdir <dir-path>
```

- Perintah untuk melihat daftar konten direktori dari *path* yang diberikan.

```
$ hadoop fs -ls
```

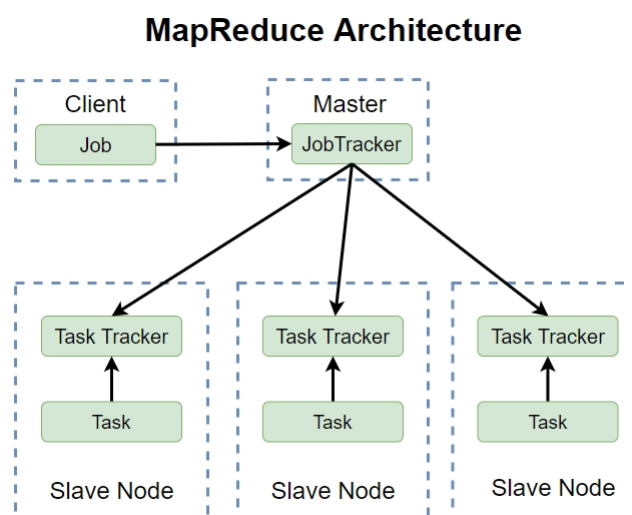
- Perintah untuk memasukkan file atau direktori lokal kepada file sistem destinasi di dalam HDFS.

```
$ hadoop fs -put <localSrc> <dest>
```

## 2.3.2 MapReduce

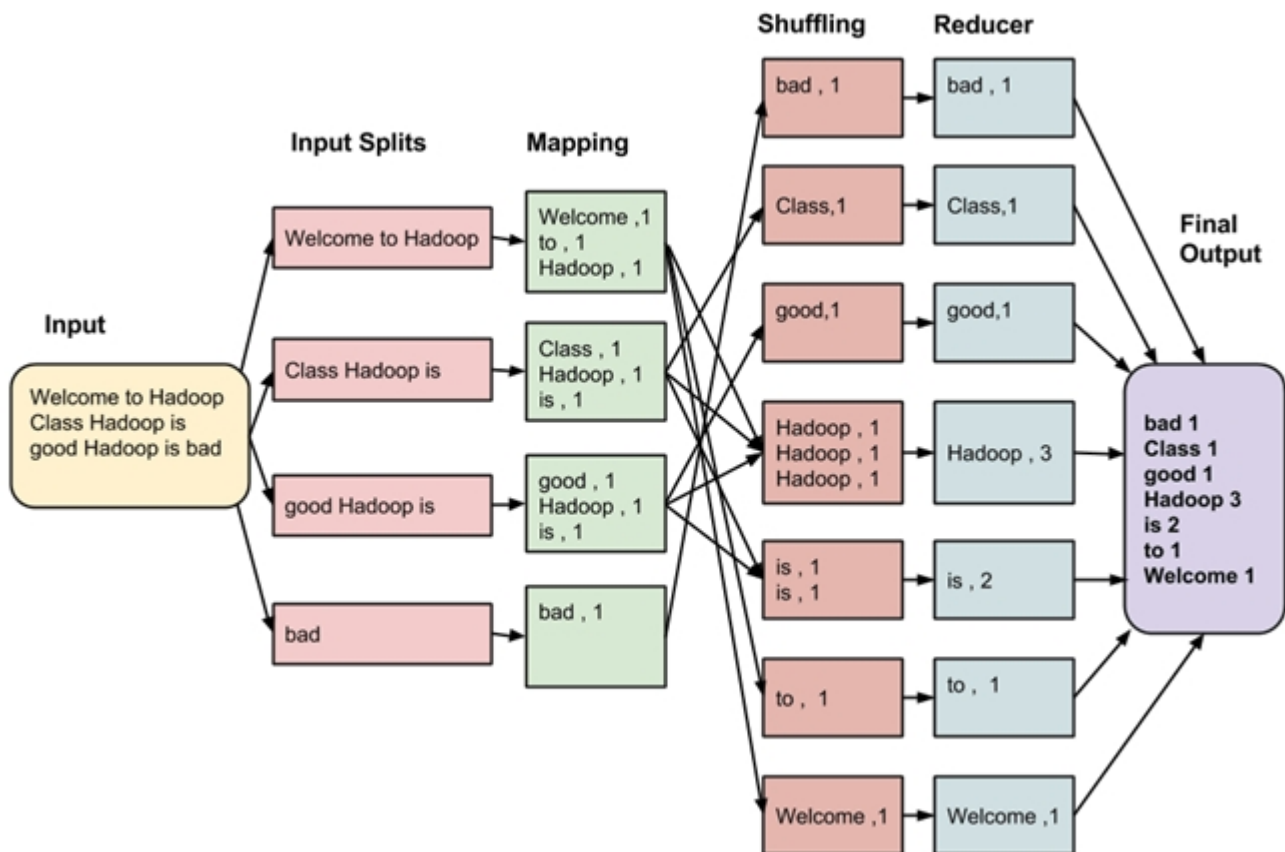
[5]

MapReduce adalah sebuah model pemrograman untuk memproses data berukuran besar secara terdistribusi dan paralel dalam *cluster* yang terdiri atas banyak komputer. Dalam memproses data, secara garis besar MapReduce dapat dibagi dalam dua proses, yaitu proses *map* dan proses *reduce*. Setiap fase memiliki pasangan *key-value* sebagai *input* dan *output*. Kedua jenis proses ini didistribusikan ke setiap komputer dalam suatu *cluster* dan berjalan secara paralel tanpa saling bergantung satu sama yang lainnya. Proses *map* bertugas untuk mengumpulkan informasi dari potongan-potongan data yang terdistribusi dalam tiap komputer dalam cluster. Hasilnya diserahkan kepada proses *reduce* untuk diproses lebih lanjut. Hasil proses *reduce* merupakan hasil akhir.



Gambar 2.15: Arsitektur MapReduce

Gambaran arsitektur MapReduce dapat dilihat pada Gambar 2.15 yaitu arsitektur MapReduce. Pada arsitektur ini, *master node* disebut JobTracker dan *slave node* disebut TaskTracker. JobTracker adalah jembatan antara pengguna dan fungsi *map* maupun *reduce*. Ketika sebuah pekerjaan *map* atau *reduce* diterima oleh JobTracker, pekerjaan tersebut akan dimasukkan ke dalam antrian. Pekerjaan dalam antrian akan dikerjakan sesuai urutan masuk pekerjaan tersebut. Kemudian, pekerjaan akan ditugaskan kepada TaskTracker oleh JobTracker. TaskTracker akan mengeksekusi pekerjaan yang diberikan oleh JobTracker dan mengembalikan laporan kemajuan kepada JobTracker.



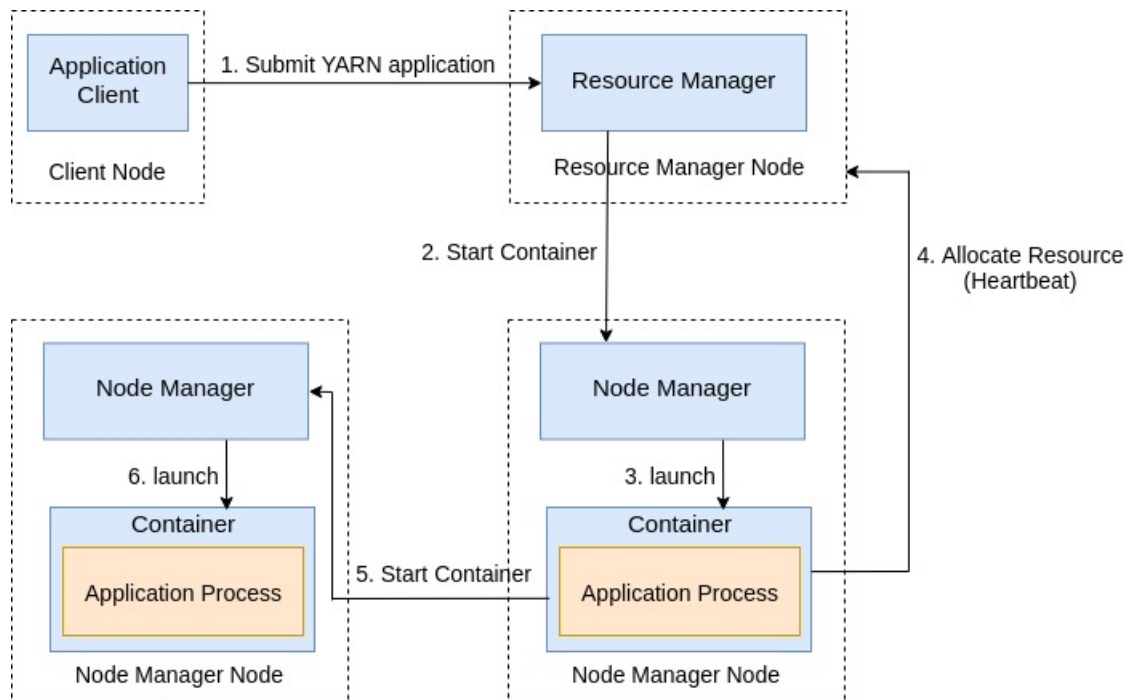
Gambar 2.16: Proses MapReduce

- Berdasarkan Gambar 2.16, berikut adalah langkah-langkah proses MapReduce:
1. *Input* dibagi menjadi *input split* yang berukuran sama. Setiap *input splits* akan dibuatkan *map task*.
  2. Pada fase *map*, data pada setiap *split* akan dihitung berapa banyak kemunculan kata tersebut dan dijadikan pasangan  $\langle \text{word}, \text{frequency} \rangle$  sebagai *output*.
  3. Fase selanjutnya adalah fase *shuffling*. Tahap ini akan mengirim *output* dari fase *map* kepada *reducer*. Hasil dari fase *map* akan dikelompokkan berdasarkan *key* dan dibagi di antara *reducer*. Dalam contoh ini, kata-kata yang sama disatukan bersama dengan frekuensi masing-masing.
  4. Terakhir adalah fase *reduce* di mana *output* dari *shuffling* akan dikumpulkan. Nilai-nilai dari fase *shuffling* akan digabungkan menjadi sebuah *output*. *Output* akan disimpan pada HDFS.



### 2.3.3 YARN

Apache YARN (*Yet Another Resource Negotiator*) adalah pengatur sumber daya dari *cluster* Hadoop. YARN bertujuan untuk memisahkan fungsionalitas antara pengaturan sumber daya dan penjadwalan pekerjaan. YARN memiliki dua tipe *daemon* yaitu *Resource Manager* dan *Node Manager* [5]. *Resource Manager* bertugas untuk mengatur sumber daya di seluruh *cluster* dan *Node Manager* yang berjalan pada *node*. *Node Manager* bertugas untuk menjalankan dan memantau *container* [5]. *Container* bertugas untuk mengeksekusi proses aplikasi yang spesifik.



Gambar 2.17: Proses menjalankan aplikasi pada YARN

Berikut adalah Gambar 2.17 yang menggambarkan langkah-langkah proses ketika menjalankan aplikasi pada YARN. Untuk menjalankan aplikasi pada YARN, *client* akan meminta *Resource Manager* untuk menjalankan proses aplikasi *master* (langkah 1). Kemudian, *Resource Manager* akan mencari *Node Manager* yang bisa menjalankan aplikasi *master* dalam sebuah *container* (langkah 2 dan 3). Ketika aplikasi *master* sudah berjalan, aplikasi *master* bisa melakukan komputasi pada *container* dan mengembalikan hasil kepada *client*. Selain itu, aplikasi *master* dapat juga meminta sumber daya tambahan (langkah 4) dan menggunakan sumber daya tersebut untuk komputasi terdistribusi (langkah 5 dan 6).

## 2.4 Spark

[7]

Apache Spark adalah sebuah *cluster computing platform* yang dirancang untuk kecepatan dan *general-purpose*. Spark dirancang berdasarkan model MapReduce yang populer untuk memberikan dukungan yang efisien kepada banyak tipe komputasi, termasuk *interactive query* dan *stream processing*. Kecepatan merupakan kunci dalam melakukan eksplorasi data. Rentang waktu dalam eksplorasi dapat dimulai dari beberapa

menit sampai beberapa jam. Salah satu fitur utama Spark yang ditawarkan adalah kemampuannya untuk melakukan *in memory computations*. Selain itu, sistem Spark lebih efisien daripada MapReduce dalam menjalankan aplikasi yang rumit pada *disk*.

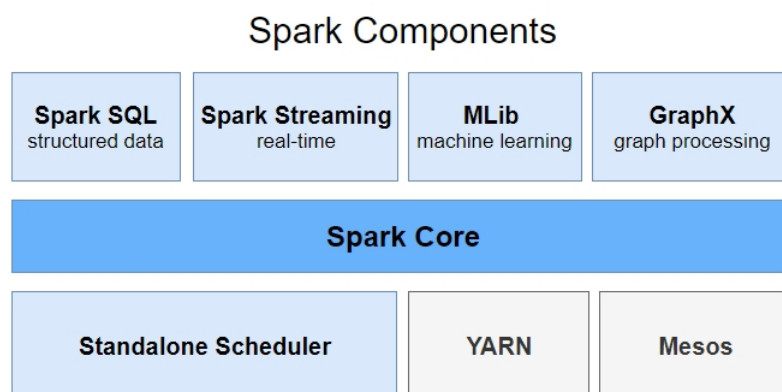
Pada sisi *general-purpose*, Spark dirancang untuk mencakup berbagai beban kerja yang sebelumnya diperlukan sistem terdistribusi terpisah, termasuk aplikasi *batch*, *iterative algorithms*, *interactive query*, dan *streaming*. Dengan mendukung beban kerja tersebut di mesin yang sama, Spark membuat pekerjaan lebih mudah dan murah untuk menggabungkan pemrosesan yang berbeda jenis. Dengan begitu, Spark mengurangi beban dalam merawat *tools* yang terpisah.

Spark dirancang untuk memudahkan pengaksesan dengan memberikan API sederhana untuk Python, Java, Scala, dan SQL. Spark dengan mudah berintegrasi dengan tools *Big Data* lainnya, terutama Hadoop. Spark bisa berjalan pada Hadoop *cluster* dan mengakses sumber data Hadoop manapun.

### 2.4.1 Komponen Spark

[7]

Spark memiliki beberapa komponen yang terintegrasi dengan erat. Sebagai *core*, Spark adalah "mesin komputasi" yang bertanggung jawab untuk penjadwalan, distribusi, dan pemantauan aplikasi yang terdiri atas banyak *task* komputasi tersebar di banyak pekerja, mesin, atau *cluster*. Karena *core engine* dari Spark sangat cepat dan dirancang untuk tujuan umum, Spark menjalankan banyak komponen di level yang lebih tinggi untuk menangani berbagai macam pekerjaan khusus seperti SQL atau *machine learning*. Komponen-komponen ini dirancang untuk saling beroperasi dengan erat. Spark mengizinkan pengguna untuk menggabungkan komponen seperti *library* dalam suatu proyek perangkat lunak.



Gambar 2.18: Komponen pada Spark

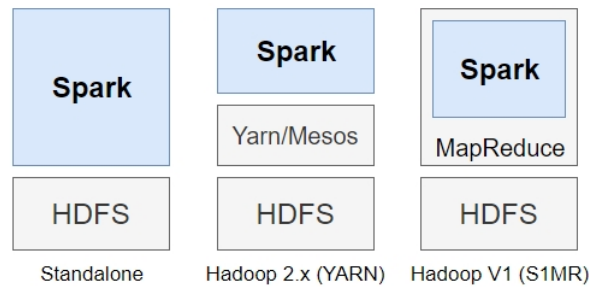
Berdasarkan Gambar 2.18, Spark memiliki beberapa komponen sebagai berikut:

- **Spark Core:** Spark Core berisi fungsi-fungsi dasar Spark, termasuk komponen untuk tugas penjadwalan, manajemen memori, pemulihan kesalahan, berinteraksi dengan sistem penyimpanan, dan banyak lagi. Spark Core memiliki banyak API *resilient distributed datasets*(RDD), yang merupakan abstraksi

1 pemrograman utama Spark. RDD mewakili suatu koleksi objek-objek yang didistribusikan di banyak  
2 node komputasi yang dapat dimanipulasi secara paralel. Spark Core menyediakan banyak API untuk  
3 membangun dan memanipulasi RDD.

- 4 • Spark SQL: Spark SQL adalah sebuah modul untuk mengerjakan data yang terstruktur. Modul ini  
5 memungkinkan melakukan *query* pada data terstruktur melalui SQL serta varian Apache Hive dari  
6 SQL yang disebut Hive Query Language (HQL) dan mendukung banyak sumber data, termasuk tabel  
7 Hive, Parquet, dan JSON. Selain menyediakan antarmuka SQL untuk Spark, Spark SQL memungkinkan  
8 *developer* untuk memadukan kueri SQL dengan fungsi-fungsi pada RDD.
- 9 • Spark Streaming: Spark Streaming adalah komponen Spark yang memungkinkan pemrosesan data  
10 dari *live streaming*. Contoh *data stream* termasuk file log yang dihasilkan oleh server web produksi,  
11 atau antrian pesan yang berisi pembaruan status yang diunggah oleh pengguna layanan web. Spark  
12 Streaming menyediakan API yang mirip dengan Spark Core's RDD API untuk memanipulasi aliran  
13 data. Hal ini membuat *developer* mudah mempelajari proyek dan berpindah antar aplikasi yang  
14 memanipulasi data yang disimpan dalam memori, pada *disk*, atau yang tiba dalam *real time*. Di balik  
15 API-nya, Spark Streaming dirancang untuk menyediakan tingkat toleransi kesalahan, *throughput*, dan  
16 skalabilitas yang sama seperti Spark Core.
- 17 • MLlib: Spark hadir dengan *library* yang berisi fungsi pembelajaran mesin (ML) secara umum, *library*  
18 ini disebut MLlib. MLlib menyediakan beberapa jenis algoritma pembelajaran mesin, termasuk  
19 klasifikasi, regresi, pengelompokan, dan penyaringan kolaboratif, serta pendukung fungsionalitas  
20 seperti *model evaluation* dan *data import*. MLlib juga menyediakan beberapa *lower-level ML primitives*,  
21 termasuk *generic gradient descent optimization algorithm*.
- 22 • GraphX: GraphX adalah sebuah *library* untuk memanipulasi grafik dan melakukan *graph-parallel*  
23 *computations*. Seperti Spark Streaming dan Spark SQL, GraphX memperluas API Spark RDD,  
24 memungkinkan pengguna untuk membuat *directed graph* dengan *arbitrary properties* yang melekat  
25 pada setiap *vertex* dan *edge*. GraphX juga menyediakan berbagai operator untuk memanipulasi grafik  
26 dan memiliki *library* yang penuh dengan *graph algorithms* yang umum seperti PageRank dan *triangle*  
27 *counting*.
- 28 • Cluster Managers: Spark dirancang untuk dapat ditambah secara efisien dari satu hingga ribuan node  
29 komputasi. Untuk mencapai hal ini dan memaksimalkan fleksibilitas, Spark dapat menjalankan lebih  
30 dari satu variasi manajer *cluster* seperti Hadoop YARN, Apache Mesos, *simple cluster manager* pada  
31 diri Spark sendiri yang disebut *Standalone Scheduler*.

### 2.4.2 Tiga Cara Membangun Spark di Atas Hadoop



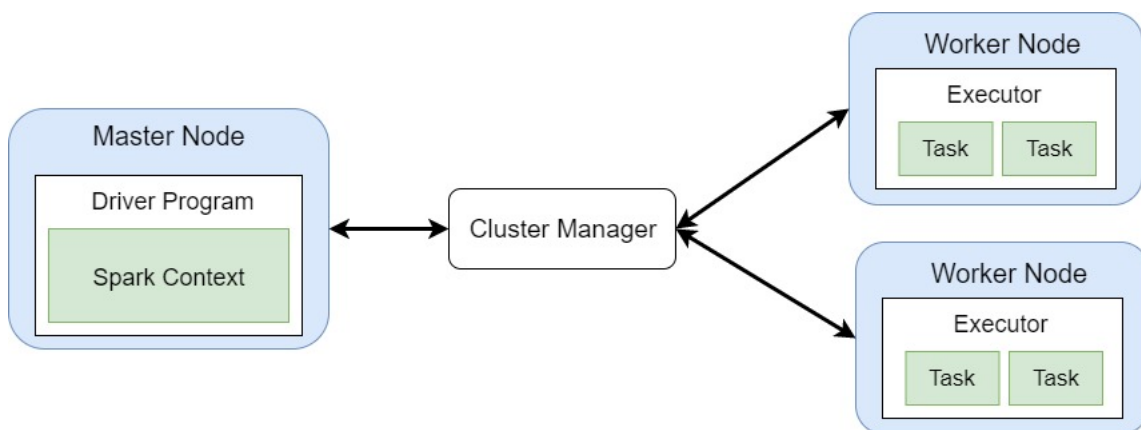
Gambar 2.19: Macam-macam cara instalasi Spark

Berdasarkan Gambar 2.19 terdapat tiga cara untuk menginstal Spark. Ketiga cara tersebut antara lain:

- *Standalone*: Spark *standalone* berarti Spark menempati tempat di atas HDFS (Hadoop Distributed File System) dan ruang dialokasikan untuk HDFS, secara eksplisit. Spark dan MapReduce akan berjalan berdampingan untuk mencakup semua pekerjaan di *cluster*.
- *Hadoop YARN*: Spark berjalan pada YARN tanpa perlu pra-instalasi atau akses root. Cara ini membantu mengintegrasikan Spark ke dalam ekosistem Hadoop. Cara ini memungkinkan komponen lain untuk berjalan di atas tumpukan.
- *Spark pada MapReduce*: Spark pada MapReduce digunakan untuk menjalankan pekerjaan-pekerjaan pada spark selain untuk *standalone deployment*. Pengguna dapat memulai Spark dan menggunakan *Spark Shell* tanpa akses administratif.

### 2.4.3 Arsitektur Spark

Spark menggunakan arsitektur *master* dan *slave*. Sebuah Spark *cluster* memiliki satu *master* dan banyak *slave* atau bisa disebut sebagai *worker*. Spark memiliki beberapa komponen penting dalam arsitekturnya seperti *Driver Program*, *Spark Context*, *Cluster Manager*. Gambar 2.20 menggambarkan komponen-komponen arsitektur Spark.



Gambar 2.20: Arsitektur Spark

Berikut adalah penjelasan dari komponen-komponen Gambar 2.20:

- **Driver Program**

*Driver program* yang berjalan pada *master node* bertugas menjalankan fungsi `main()` dari aplikasi dan tempat di mana *Spark Context* dibuat. Kode program akan diterjemahkan menjadi *tasks* dan dijadwalkan kepada *executors* untuk dikerjakan. *Driver program* akan berkomunikasi dengan *cluster manager* untuk mengatur sumber daya pada *cluster*.

- **Spark Context**

*Spark Context* menghubungkan pengguna dengan *cluster*. *Spark Context* dapat terhubung dengan beberapa *cluster manager* seperti YARN, MESOS, dan *Spark standalone cluster manager*. *Spark Context* dapat digunakan untuk membuat *Resilient Distributed Datasets* (RDD), *accumulators*, dan *broadcast variable*.

- **Cluster Manager**

*Cluster Manager* berfungsi mengatur sumber daya pada sebuah *cluster*. Spark dapat berjalan pada berbagai macam *cluster manager* seperti Apache Mesos, Hadoop YARN, dan *Spark's stand alone*. *Cluster manager* akan berusaha mendapatkan sumber daya pada *cluster* dan mengalokasikannya kepada *Spark job* yang sedang berjalan.

- **Executors**

*Executors* adalah proses-proses yang berjalan pada *worker node* dan bertanggung jawab untuk mengerjakan *tasks* yang diberikan. *Executors* dibuat ketika aplikasi dijalankan dan akan tetap ada selama aplikasi masih berjalan.

- **Tasks**

*Task* adalah sebuah satuan kerja pada Spark. *Task* berisi perintah-perintah. Perintah tersebut merupakan fungsi yang diserialisasi. *Task* akan dikirimkan oleh *driver program* kepada *executor*. Kemudian, *executor* akan mendeserialisasi perintah tersebut dan mengerjakannya. Pada umumnya *task* akan dibuat untuk setiap partisi. Partisi merupakan potongan data yang terdistribusi pada *cluster*.

#### 2.4.4 Resilient Distributed Datasets (RDD)

*Resilient Distributed Datasets* (RDD) adalah struktur data dasar pada Spark yang berisi koleksi benda-benda yang didistribusikan secara permanen. Setiap dataset dalam RDD dibagi menjadi beberapa partisi yang dapat dikomputasi pada node yang berbeda pada *cluster* [7]. RDD dapat berisi jenis objek Python, Java, atau Scala, termasuk kelas yang ditentukan pengguna. Spark memanfaatkan konsep RDD untuk mencapai operasi MapReduce yang lebih cepat dan efisien [7].

Secara umum, RDD merupakan kumpulan *read-only, partitioned collection* dari *records*. RDD dapat dibuat melalui operasi deterministik dari data pada penyimpanan yang stabil atau RDD lainnya [7]. Elemen pada RDD memiliki sifat *fault tolerance* dan dapat dioperasikan secara paralel.

*Data sharing* pada MapReduce lebih lambat dibanding RDD karena replikasi, serialisasi, dan disk IO. Sebagian besar aplikasi Hadoop menghabiskan lebih dari 90 persen waktunya untuk melakukan operasi *read-write* kepada HDFS.

Untuk menangani masalah tersebut, dibangun *framework* khusus yang disebut Apache Spark. Ide utama dari Spark adalah RDD, Spark juga mendukung *in-memory computation*. Spark menyimpan status memori sebagai objek di seluruh pekerjaan dan objek dapat dibagi di antara *jobs*. *Data sharing* dalam memori lebih cepat 10 hingga 100 kali lipat dibanding *network* atau *disk*.

Berikut adalah sifat-sifat dari RDD [7]:

- *In Memory*: Data pada RDD disimpan pada memori sebesar mungkin dan selama mungkin.
- *Partitioned*: *records* dipartisi dan didistribusikan kepada *node-node* di dalam *cluster*.
- *Typed*: RDD memiliki tipe data seperti RDD[Long], RDD[String] dan tipe data lainnya.
- *Lazy evaluation*: Data di dalam RDD tidak akan tersedia atau berubah sampai sebuah perintah *action* telah dieksekusi.
- *Immutable*: RDD yang telah dibuat tidak dapat berubah. Meskipun demikian, RDD dapat ditransformasi menjadi sebuah RDD baru dengan melakukan perintah *transformation* pada RDD.
- *Parallel*: RDD dapat dioperasikan secara paralel.
- *Cacheable*: Pengguna dapat memilih RDD mana yang akan dipakai kembali dan memilih tempat penyimpanannya, yaitu memori atau *disk*. Dengan begitu, data dapat diakses lebih cepat untuk permintaan selanjutnya.

Terdapat dua cara untuk membuat sebuah RDD. Cara pertama adalah dengan memuat dataset eksternal, sedangkan cara alternatif adalah dengan mendistribusikan sebuah koleksi objek seperti *list* atau *set* [7]. Terdapat dua tipe operasi yang dapat dilakukan RDD, yaitu *transformations* dan *actions*. *Transformations* membuat RDD baru dari RDD sebelumnya [7]. Berbeda dengan *transformations*, *actions* mengembalikan nilai hasil komputasi berdasarkan RDD [7]. Hasil dari *actions* akan dikembalikan kepada *driver program* atau disimpan pada penyimpanan eksternal seperti HDFS.

Berikut adalah contoh pembuatan RDD dari sumber eksternal dan koleksi objek:

- `val lines = sc.textFile("/path/to/README.md")` \\ sumber eksternal
- `val lines = sc.parallelize(["a", "b", "c", "d", "e"])` \\ array

*Transformations* pada RDD adalah sebuah operasi yang menerima RDD sebagai masukan dan mengembalikan satu atau lebih RDD baru. RDD masukan tidak berubah karena sifat RDD adalah *immutable* yang berarti tidak bisa diubah ketika dibuat. *Transformations* bersifat *lazy* dan tidak langsung dieksekusi, Spark akan mencatat *transformation* apa saja yang dilakukan pada RDD sejak awal. *Transformations* akan dieksekusi ketika sebuah *actions* dipanggil.

Berikut adalah contoh *filter transformation* di Scala. *Filter* digunakan untuk menyaring elemen-elemen yang sesuai dengan kriteria yang ditentukan. Pada kasus ini, filter akan mengambil baris-baris yang memiliki kata *error*.

```

1 val inputRDD = sc.textFile("log.txt")
2 val errorsRDD = inputRDD.filter(line => line.contains("error"))

```

3     Tabel 2.4 berisi daftar *transformations* yang umum pada Spark:

Tabel 2.4: Tabel transformations

<i>Transformations</i>	Penjelasan
<b>map</b> (func)	Mengembalikan RDD baru yang dibentuk dengan melewati setiap elemen melalui fungsi func.
<b>mapPartitions</b> (func)	Mengembalikan RDD baru yang dibentuk dengan melewati setiap partisi melalui fungsi func.
<b>filter</b> (func)	Mengembalikan RDD baru yang dibentuk dengan memilih elemen-elemen yang mengembalikan nilai <i>true</i> dari fungsi func.
<b>flatMap</b> (func)	Mirip dengan <i>map</i> , tetapi setiap elemen dapat dipetakan menjadi nol atau lebih elemen sebagai keluaran.
<b>union</b> (otherDataset)	Mengembalikan RDD baru yang mengandung elemen dari kedua sumber.
<b>intersection</b> (otherDataset)	Mengembalikan RDD baru yang berisi potongan elemen dari sumber dan sumber lainnya.
<b>distinct</b> ([numPartitions])	Mengembalikan RDD baru yang mengandung elemen yang unik dari sumber.
<b>groupByKey</b> ([numPartitions])	Mengembalikan RDD baru bertipe <i>pairs</i> (K, Iterable<V>) dari sumber RDD bertipe (K, V).
<b>groupByKey</b> (func,[numPartitions])	Mengembalikan RDD baru berupa <i>pairs</i> (K, V) yang sudah diagregasi berdasarkan <i>key</i> dan fungsi <i>reduce</i> yang diberikan.
<b>sortByKey</b> ([ascending], [numPartitions])	Mengembalikan RDD baru berupa <i>pairs</i> (K, V) yang terurut secara menaik atau menurun berdasarkan parameter boolean yang diberikan.
<b>join</b> (otherDataset, [numPartitions])	Mengembalikan gabungan RDD berupa <i>pairs</i> (K, V) dan (K, W) menjadi <i>pairs</i> (K, (V,W)).

4     Berikut adalah contoh operasi *action* pada RDD. Pada contoh ini, fungsi *reduceByKey* digunakan untuk  
5 menghitung jumlah kata yang ada.

```

6     val lines = sc.textFile("data.txt")
7     val pairs = lines.map(s => (s, 1))
8     val counts = pairs.reduceByKey((a, b) => a + b)

```

9     *Actions* merupakan operasi yang mengembalikan sebuah nilai kepada *driver program* atau tempat penyimpanan eksternal. Untuk mengembalikan sebuah nilai, dapat digunakan fungsi-fungsi seperti *take()*, *count()*, *collect()*, dan *actions* lainnya. Operasi *take()* digunakan untuk mengambil sebagian kecil elemen pada RDD. Ketika menggunakan *collect()*, memori pada satu komputer harus cukup untuk menampung seluruh *data set* [7]. Operasi tersebut sebaiknya digunakan pada *data set* yang berukuran kecil. *Data set* yang berukuran besar dapat disimpan pada tempat penyimpanan eksternal. Setiap kali sebuah *actions* dipanggil, seluruh RDD akan dikomputasi dari akarnya. Untuk mencapai efisiensi yang lebih tinggi, dapat dilakukan *persist* terhadap *intermediate results*.



Berikut adalah Tabel 2.5 berisi daftar *actions* yang umum pada Spark:

Tabel 2.5: Tabel Actions

<i>Actions</i>	Penjelasan
<b>reduce(func)</b>	Mengagregasikan seluruh elemen pada RDD menggunakan fungsi yang diberikan pada <i>parameter</i> .
<b>collect()</b>	Mengembalikan seluruh <i>data set</i> sebagai <i>array</i> kepada <i>driver program</i> .
<b>count()</b>	Mengembalikan jumlah elemen pada RDD.
<b>first()</b>	Mengembalikan elemen pertama pada RDD.
<b>take(n)</b>	Mengembalikan sebuah <i>array</i> dengan n jumlah elemen pertama dari RDD.
<b>takeOrdered(n, [ordering])</b>	Mengembalikan sebuah <i>array</i> dengan n jumlah elemen pertama dari RDD secara terurut.
<b>saveAsTextFile(path)</b>	Menyimpan <i>dataset</i> sebagai <i>text file</i> pada direktori yang ditentukan.
<b>saveAsSequenceFile(path)</b>	Menyimpan RDD sebagai Hadoop SequenceFile pada direktori yang ditentukan.
<b>saveAsObjectFile(path)</b>	Menyimpan RDD sebagai format yang sederhana menggunakan Java Serialization pada direktori yang ditentukan.
<b>countByKey()</b>	Menjumlahkan <i>pairs</i> (K, V) berdasarkan <i>key</i> dan mengembalikan sebuah <i>pairs</i> berisi (K, int).
<b>foreach(func)</b>	Memproses setiap elemen pada RDD menggunakan fungsi <i>func</i> yang diberikan.

## 2.5 Scala

Scala adalah sebuah bahasa pemrograman yang diciptakan oleh Martin Odersky, yaitu seorang Profesor di Ecole Polytechnique Federale de Lausanne, sebuah kampus di Lausanne, Swiss. Kata Scala sendiri merupakan singkatan dari "Scalable Language". Karena Scala berjalan di atas *Java Virtual Machine* (JVM), Scala memiliki performa yang relatif cepat dan juga memungkinkan untuk menggabungkan kode di Scala dengan di Java. library, framework dan tool yang ada di Java dapat digunakan pada Scala. Scala menggabungkan konsep *Object Oriented Programming* (OOP) yang dikenal di Java dengan konsep *Functional Programming* (FP). Adanya konsep FP inilah yang menjadikan Scala sangat ekspresif, nyaman dan menyenangkan untuk digunakan.

Perintah *scalac* digunakan untuk mengkompilasi program Scala dan akan menghasilkan beberapa file kelas di direktori saat ini. Salah satunya akan disebut file *.class*. Ini adalah *bytecode* yang akan berjalan di JVM dengan menggunakan perintah *scala*.

### 2.5.1 Expressions

*Expressions* adalah pernyataan atau argumen yang dapat dikomputasi.

$1 + 1$



```
1 2 + 2
```

2 *Expressions* dapat dikembalikan dengan perintah *println*.

```
3 println(1)
4 println(100) // 100
5 println(1 + 1) // 2
6 println("Hi!") // Hi!
```

7 *Expressions* atau pernyataan seperti di atas dapat disimpan dalam sebuah *variable*. Terdapat dua jenis  
8 *variable* di Scala yaitu *val* dan *var*. Setelah *val* diinisialisasi, *val* tidak dapat diisi kembali yang berarti nilai  
9 dari *val* tidak dapat diubah.

```
10 val x = 2 + 5
11 val x = 10 //tidak akan di-compile
12 val y = 7
13 val coba:Int = 200
```

14 *variable* mirip dengan value, tetapi nilai *variable* dapat diisi kembali.

```
15 var x = 2 + 2
16 x = 4
17 println(x) // 4
18 x = 7
19 println(x) // 7
```

20 Secara eksplisit, *developer* dapat menyatakan tipe dari sebuah *var* atau *val* dengan cara:

```
21 var x: Int = 1 + 1 // Int merupakan tipe dari variable x
22 val y: Long = 987654321 // Long merupakan tipe dari variable y
23 val z: Char = 'a' // Char merupakan tipe dari variable z
```

## 24 2.5.2 Blocks

25 *Block* digunakan untuk menggabungkan *expressions*. Berikut adalah contoh *blok*:

```
26 println({
27     val x = 1 + 1
28     x + 1
29 }) // 3
```

## 30 2.5.3 Loop dan Conditional

31 *loop* merupakan struktur pengulangan yang memungkinkan menulis suatu *loop* yang perlu dieksekusi sekian  
32 kali secara efisien. Terdapat berbagai bentuk *loop* dalam Scala yang dijelaskan di bawah ini:

```
1  for( var x <- Range ){
2      statement(s);
3  }
4
5  var x = 0
6  while (x < 10) {
7      println(x)
8      x += 1
9  }
```

10     *COnditional* atau percabangan adalah pengujian sebuah kondisi. Jika kondisi yang diuji tersebut terpenuhi,  
11     maka program akan menjalankan pernyataan-pernyataan tertentu. Jika kondisi yang diuji salah, program  
12     akan menjalankan pernyataan yang lain. Berikut adalah contoh percabangan dalam bahasa Scala:

```
13  if( x < 20 ){
14      println("This is if statement");
15  }
16
17  if( x < 20 ){
18      if( x < 5) {
19          println("smallest");
20      }
21  }
22
23  if( x < 10 ){
24      println("This is bigger");
25  } else {
26      println("This is smaller");
27  }
28
29  if( x == 1 ){
30      println("1");
31  } else if (x == 2){
32      println("2");
33  }
```

#### 34   **2.5.4 Functions**

35     *Functions* adalah *expression* yang mempunyai atau menerima parameter. Sebuah *function* yang tidak memiliki  
36     nama disebut *anonymous function*. Berikut adalah contoh *anonymous function* dan *function* biasa. Sebuah  
37     *function* dapat memiliki lebih dari satu parameter.

```
38  (x: Int) => x + 1 // Anonymous function
```

39

```
1 val addOne = (x: Int) => x + 1 // function biasa
2 println(addOne(2)) // 3
3
4 val add = (x: Int, y: Int) => x + y
5 println(add(1, 2)) // 3
```

6 Pada sisi sebelah kiri tanda "=>" adalah parameter-parameter sebuah *function*, sementara pada sisi  
7 sebelah kanan merupakan ekspresi-ekspresi yang melibatkan parameter tersebut.  
8

### 9 2.5.5 Methods

10 *Method* sangat mirip dengan *function*, tetapi *method* memiliki beberapa perbedaan. *Method* harus didefini-  
11 sikan dengan kata kunci *def*, diikuti dengan nama *method*, parameter-parameter dari *method* tersebut, tipe  
12 kembalian *method*, dan isi dari *method* tersebut.

```
13 def add(x: Int, y: Int): Int = x + y
14 println(add(1, 2)) // 3
```

15 *Method* dapat mempunyai lebih dari satu parameter.

```
16 def addThenMultiply(x: Int, y: Int)(multiplier: Int): Int = (x + y) * multiplier
17 println(addThenMultiply(1, 2)(3)) // 9
```

18 *Method* dapat tidak memiliki parameter.

```
19 def name: String = System.getProperty("user.name")
20 println("Hello, " + name + "!")
```

21 *Method* berbeda dengan *functions* dapat memiliki *multi-line expressions*

```
22 def getSquareString(input: Double): String = {
23     val square = input * input
24     square.toString
25 }
```

26 *Expression* terakhir dari *method* menjadi nilai yang akan dikembalikan. Scala mempunyai *keyword* *return*,  
27 tetapi sangat jarang digunakan.

### 28 2.5.6 Class dan Object

29 *Class* pada Scala didefinisikan dengan kata kunci *class* yang diikuti dengan namanya dan terakhir adalah  
30 *constructor* parameter.

```
31 class Greeter(prefix: String, suffix: String) {
32     def greet(name: String): Unit = {
33         println(prefix + name + suffix)
```

```
1      }  
2  }  
3
```

4 Berikut adalah cara mendeklarasi sebuah objek pada Scala

```
5 val greeter = new Greeter("Hello, ", "!")  
6 greeter.greet("Scala developer")
```

7 Objek dapat dianggap sebagai suatu instansi tunggal pada kelas itu sendiri. Kata kunci *object* dapat  
8 digunakan untuk mendefinisikan sebuah objek.

```
9 object IdFactory {  
10     private var counter = 0  
11     Main method  
12     def create(): Int = {  
13         counter += 1  
14         counter  
15     }  
16 }  
17  
18 val newId: Int = IdFactory.create()  
19 println(newId) // 1  
20 val newerId: Int = IdFactory.create()  
21 println(newerId) // 2  
22
```

23 *Main method* adalah pintu masuk dari sebuah program. JVM membutuhkan sebuah *main method* yang  
24 dinamakan *main* dan menerima satu *argument*, yaitu sebuah *array* bertipe *string*. Menggunakan *object*,  
25 *developer* dapat mendefinisikan sebuah *main method* seperti berikut:

```
26 object Main {  
27     def main(args: Array[String]): Unit = {  
28         println("Hello, Scala developer!")  
29     }  
30 }
```

### 31 2.5.7 Higher Order Function

32 Pada bahasa Scala, terdapat sebuah fungsi yang disebut sebagai *Higher Order Function*. *higher order function*  
33 merupakan sebuah fungsi yang menerima fungsi lainnya sebagai *parameter* dan mengembalikan sebuah fungsi  
34 sebagai hasilnya. Berikut adalah contoh-contoh *higher order function*:

```
35 val salaries = Seq(20000, 70000, 40000)  
36 val doubleSalary = (x: Int) => x * 2  
37 val newSalaries = salaries.map(doubleSalary) // List(40000, 140000, 80000)  
38
```

1 Kode program dapat dipersingkat dengan menggunakan fungsi *anonymous* dan langsung dimasukkan  
2 pada *parameter*.

```
3 val salaries = Seq(20000, 70000, 40000)
4 val newSalaries = salaries.map(x => x * 2) // List(40000, 140000, 80000)
```

5 *Developer* juga dapat memasukkan *method* pada *parameter higher order function*, *compiler* Scala akan  
6 mengubah sebuah *method* menjadi fungsi.

```
7 case class WeeklyWeatherForecast(temperatures: Seq[Double]) {
8
9     private def convertCtoF(temp: Double) = temp * 1.8 + 32
10
11     def forecastInFahrenheit: Seq[Double] = temperatures.map(convertCtoF)
12 }
```

13 Salah satu alasan untuk menggunakan *higher order function* adalah untuk mengurangi kode yang  
14 berlebihan. Misalkan terdapat beberapa metode yang dapat menaikkan gaji seseorang dengan berbagai faktor.  
15 Tanpa membuat *higher order function*, kode akan terlihat seperti berikut:

```
16 object SalaryRaiser {
17
18     def smallPromotion(salaries: List[Double]): List[Double] =
19         salaries.map(salary => salary * 1.1)
20
21     def greatPromotion(salaries: List[Double]): List[Double] =
22         salaries.map(salary => salary * math.log(salary))
23
24     def hugePromotion(salaries: List[Double]): List[Double] =
25         salaries.map(salary => salary * salary)
26 }
```

27 Perhatikan bahwa masing-masing dari ketiga *method* hanya berbeda pada faktor perkalian. Untuk  
28 menyederhanakan kode tersebut, *developer* dapat mengeluarkan kode yang redundan menjadi *higher order*  
29 *function* seperti:

```
30 object SalaryRaiser {
31
32     private def promotion(salaries: List[Double], promoF: Double => Double): List[Double] =
33         salaries.map(promotionFunction)
34
35     def smallPromotion(salaries: List[Double]): List[Double] =
36         promotion(salaries, salary => salary * 1.1)
37
38     def bigPromotion(salaries: List[Double]): List[Double] =
```

```
1      promotion(salaries, salary => salary * math.log(salary))
2
3  def hugePromotion(salaries: List[Double]): List[Double] =
4      promotion(salaries, salary => salary * salary)
5  }
```

## BAB 3

### STUDI DAN EKSPLORASI APACHE SPARK

Pada bab ini, akan dijelaskan eksplorasi yang dilakukan pada Spark. Studi dan eksplorasi dilakukan untuk mengetahui lebih tentang fungsi-fungsi RDD pada Spark, cara instalasi, Spark *shell*, dan Spark UI.

#### 3.1 Instalasi Apache Spark

Berikut adalah tahap-tahap untuk melakukan instalasi Apache Spark. Apache Spark yang digunakan adalah Apache Spark versi 2.3.1. Spark dapat berjalan di atas berbagai sistem operasi seperti Windows dan UNIX systems (Contoh Linux, macOS). Sebelum memulai instalasi Apache Spark, terdapat beberapa kebutuhan yang harus dipenuhi seperti instalasi Java dan Scala. Berikut adalah langkah-langkah untuk memastikan bahwa kebutuhan minimal telah terpenuhi:

- Pastikan bahwa Java telah diinstal dan versi java yang diinstall adalah setidaknya 8+ karena Spark berjalan pada versi minimal Java 8+. Berikut adalah command untuk memastikan java telah terinstall:

```
$ java -version
Java(TM) SE Runtime Environment (build 1.8.0_112-b15)
```

- Pastikan bahwa Scala telah diinstal dengan versi minimal 2.11.x. Berikut adalah perintah untuk memastikan bahwa Scala telah terinstal dengan versi yang benar:

```
$ scala -version
Scala code runner version 2.11.6 -- Copyright 2002-2013, LAMP/EPFL
```

Bila Java dan Scala belum terinstal pada komputer, berikut adalah langkah-langkah instalasi Java dan Scala untuk kebutuhan Spark:

- Berikut adalah perintah-perintah untuk menginstal Java menggunakan terminal pada sistem operasi Linux:

```
$ sudo apt-get update
$ sudo apt-get install default-jdk
```

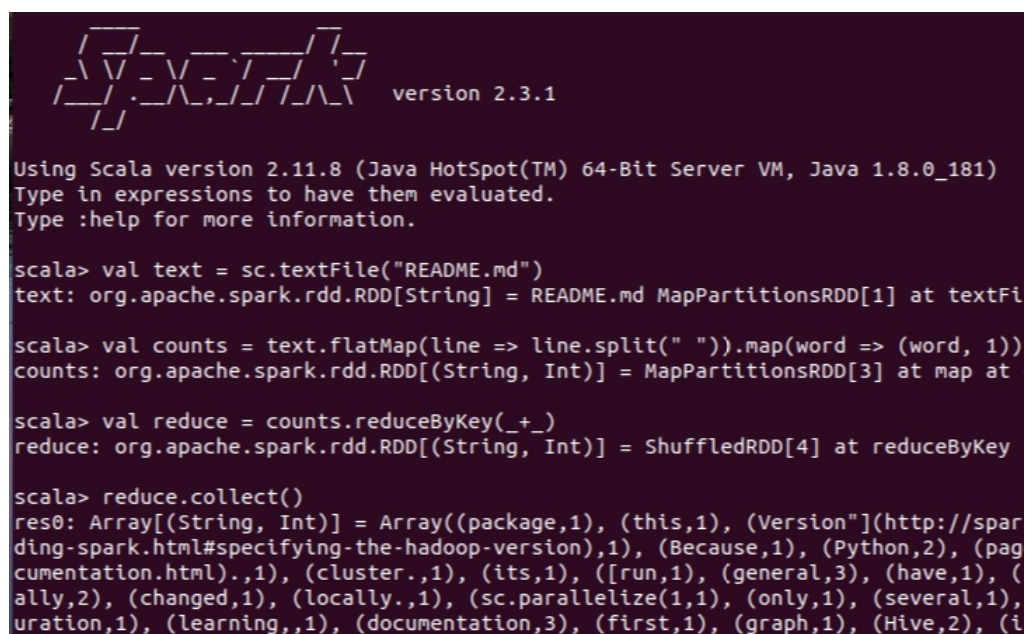
- Berikut adalah perintah-perintah untuk menginstal Scala menggunakan terminal pada sistem operasi Linux:





## 3.2 Eksplorasi Spark Shell

Bagian ini menjelaskan percobaan untuk menghitung jumlah setiap kata pada file *text* README.md. Spark *shell* digunakan untuk menjalankan perintah-perintah agar Spark bisa menghitung jumlah setiap kata yang ada pada file *text* tersebut. Setiap kata yang sama akan dijumlahkan. Pada bagian ini akan digunakan *transformation* dan juga *action*.



```

version 2.3.1

Using Scala version 2.11.8 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_181)
Type in expressions to have them evaluated.
Type :help for more information.

scala> val text = sc.textFile("README.md")
text: org.apache.spark.rdd.RDD[String] = README.md MapPartitionsRDD[1] at textFi

scala> val counts = text.flatMap(line => line.split(" ")).map(word => (word, 1))
counts: org.apache.spark.rdd.RDD[(String, Int)] = MapPartitionsRDD[3] at map at

scala> val reduce = counts.reduceByKey(_+_ )
reduce: org.apache.spark.rdd.RDD[(String, Int)] = ShuffledRDD[4] at reduceByKey

scala> reduce.collect()
res0: Array[(String, Int)] = Array((package,1), (this,1), (Version,1)(http://spar
ding-spark.html#specifying-the-hadoop-version),1), (Because,1), (Python,2), (pag
cumentation.html),1), (cluster.,1), (its,1), ([run,1), (general,3), (have,1), (
ally,2), (changed,1), (locally.,1), (sc.parallelize(1,1), (only,1), (several,1),
uration,1), (learning,,1), (documentation,3), (first,1), (graph,1), (Hive,2), (i

```

Gambar 3.2: Word Count

Berdasarkan Gambar 3.2, berikut adalah langkah-langkah percobaan yang dilakukan:

1. Jalankan spark shell dengan *command* berikut pada terminal:

```
$ ./bin/spark-shell
```

2. Buat *text* RDD dari sumber eksternal, yaitu file README.md. *Command* di bawah digunakan untuk membuat RDD dari file eksternal:

```
scala> val text = sc.textFile("README.md")
```

Dapat dilihat bahwa RDD bertipe *String* telah sukses dibuat.

```
text: org.apache.spark.rdd.RDD[String] = README.md MapPartititonsRDD[1]...
```

3. Gunakan operasi *transformation* `flatMap()` untuk memecah kalimat menjadi kata-kata. Setelah itu, setiap kata akan dijadikan pasangan *key* (kata) dan *value* (kata,1). Berikut adalah perintah yang harus dijalankan:

```
val counts = text.textflatMap(line => line.split(" ")).map(word => (word, 1))
counts: org.apache.spark.rdd.RDD[(String, int)] = ShuffledRDD[3] ...
```

4. Hitung jumlah setiap kata dengan menggunakan operasi `reduceByKey()`. Operasi `reduceByKey()` akan menjumlahkan kata dengan *key* yang sama. Contoh perintah dapat dilihat dibawah:

```
val reduce = counts.reduceByKey(_+_)  
reduce: org.apache.spark.rdd.RDD[(String, int)] = ShuffledRDD[4] ...
```

5. Ambil hasil operasi sebelumnya dengan menggunakan operasi `collect()` yang merupakan sebuah *action*. Berikut adalah perintah yang harus dijalankan:

```
reduce.collect()  
//Hasil  
res0: Array[(String, Int)] = Array((package,1), (Python,2), .....
```

### 3.3 Instalasi Apache Spark pada *Multi-Node Cluster*

Seperti yang telah disebutkan sebelumnya, Apache Spark dapat diterapkan *multi-node cluster*. Berikut adalah langkah-langkah yang harus dilakukan:

1. Tambahkan entri dalam file host *master* dan *slave*. *Master* merupakan komputer utama dan *slave* merupakan komputer pekerja. Berikut adalah perintah yang harus dijalankan:

```
$ sudo gedit /etc/hosts
```

Tambahkan IP *master* dan juga *slave* pada file.

```
<MASTER-IP> master  
<SLAVE1-IP> slave1  
<SLAVE2-IP> slave2  
<SLAVE3-IP> slave3
```

2. Install Java pada setiap *master* dan *slave*, jangan lupa untuk memastikan versi Java yang di install. Berikut adalah perintah untuk menginstal Java:

```
$ sudo apt-get update  
$ sudo apt-get install default-jdk
```

Pastikan versi Java yang diinstal dengan perintah berikut:

```
$ java -version
```

3. instal Scala pada setiap master dan slave, jangan lupa untuk memastikan versi Scala yang diinstal.

```
$ sudo apt-get update  
$ sudo apt-get install scala
```

Pastikan veri Scala yang diinstal dengan perintah berikut:

```
$ scala -version
```

4. Setelah melakukan instalasi Scala dan Java, Instal Open SSH Server-Client pada *master*. Berikut adalah perintah yang harus dijalankan:

```
$ sudo apt-get install openssh-server openssh-client
```

```
$ ssh-keygen -t rsa -P
```

5. Lakukan konfigurasi SSH pada *slave* dan juga *master*. Salin `.ssh/id_rsa.pub` milik *master* kepada `.ssh/authorized_keys` untuk *master* dan juga *slave*.

6. Setelah itu, kita akan mengunduh dan menginstal Spark pada setiap *slave* dan *master*. Berikut adalah langkah-langkah yang diikuti:

Unduh versi Spark yang diinginkan pada <https://spark.apache.org/downloads.html>

Ekstrak Spark dengan perintah berikut:

```
$ tar xvf spark-2.3.0-bin-hadoop2.7.tgz
```

```
$ sudo mv spark-2.3.0-bin-hadoop2.7 /home/user/spark
```

7. Setelah selesai menginstal Spark, kita harus mengubah file `.bashrc`. Buka file `bashrc` dengan command berikut:

```
$ sudo gedit .bashrc
```

Tambahkan baris berikut pada file `.bashrc`:

```
export PATH = $PATH:/home/user/spark/bin
```

Jalankan perintah berikut untuk memastikan perubahan telah terjadi pada file `.bashrc`:

```
source .bashrc
```

8. Lakukan konfigurasi pada *master* dengan mengubah file `spark-env.sh`. Berikut adalah perintah-perintah yang harus dijalankan

```
$ cd /home/user/spark/conf
```

```
$ cp spark-env.sh.template spark-env.sh
```

```
$ sudo gedit spark-env.sh
```

```
source .bashrc
```

Tambahkan baris berikut pada file tersebut:

```
1 export SPARK_MASTER_HOST='<MASTER-IP>'
2 export JAVA_HOME=<Path_of_JAVA_installation>
```

3 Kemudian edit file slaves pada /home/user/spark/conf dengan perintah berikut:

```
4 $ sudo gedit slaves
```

5 Tambahkan baris berikut pada file tersebut:

```
6 master
7 slave1
8 slave2
9 slave3
```

10 9. Jalankan spark *cluster* dengan perintah berikut:

```
11 $ cd /usr/local/spark
12 $ ./sbin/start-all.sh
```

13 Untuk memberhentikannya masukan perintah berikut:

```
14 $ ./sbin/start-all.sh
```

### 15 3.4 Percobaan Spark Submit

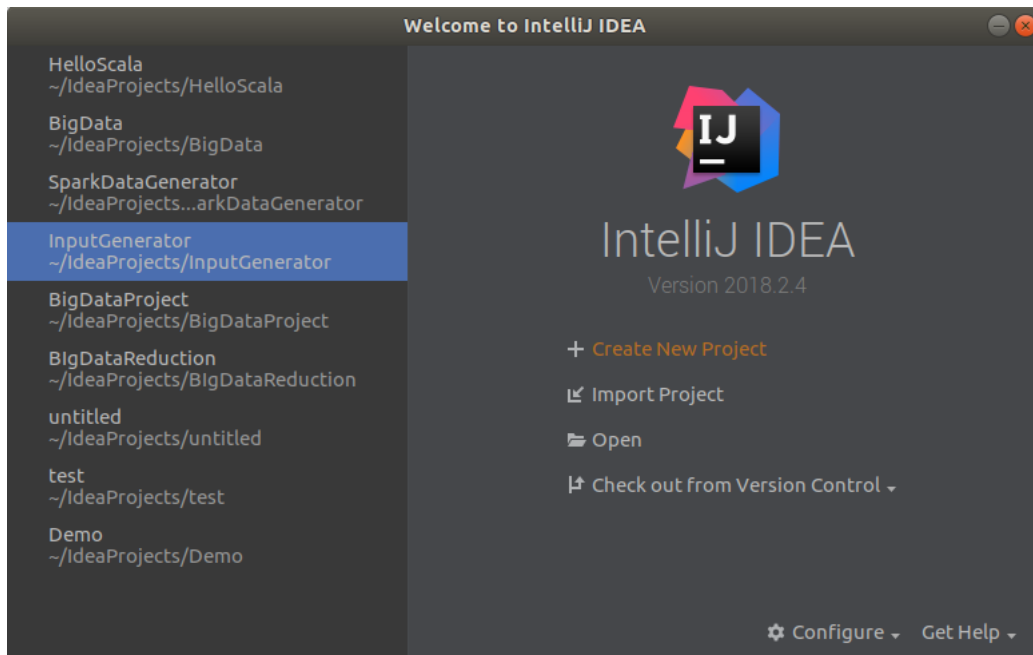
16 Pada percobaan ini, kita akan mencoba mengumpulkan sebuah jar kepada spark-submit. Aplikasi yang dibuat  
17 harus memiliki konfigurasi Spark dan diubah menjadi jar untuk dikumpulkan kepada spark-submit. Aplikasi  
18 yang dibuat akan membaca file yang disediakan dan menghitung jumlah kata yang ada. Sebelum melakukan  
19 percobaan, terdapat beberapa kebutuhan yang harus dipenuhi. Berikut adalah kebutuhan-kebutuhan yang  
20 harus dipenuhi:

- 21 1. Instal dan sudah melakukan konfigurasi untuk Scala, Java, dan Spark.
- 22 2. Instal IntelliJ IDEA dari <https://www.jetbrains.com/idea/>.
- 23 3. Install sbt, berikut adalah langkah instalasi sbt:

```
24 $ echo "deb https://dl.bintray.com/sbt/debian /" | sudo tee -a /etc/apt/sources.list.d/sb
25 $ sudo apt-key adv --keyserver hkp://keyserver.ubuntu.com:80 --recv 2EE0EA64E40A89B84B2DF
26 $ sudo apt-get update
27 $ sudo apt-get install sbt
```

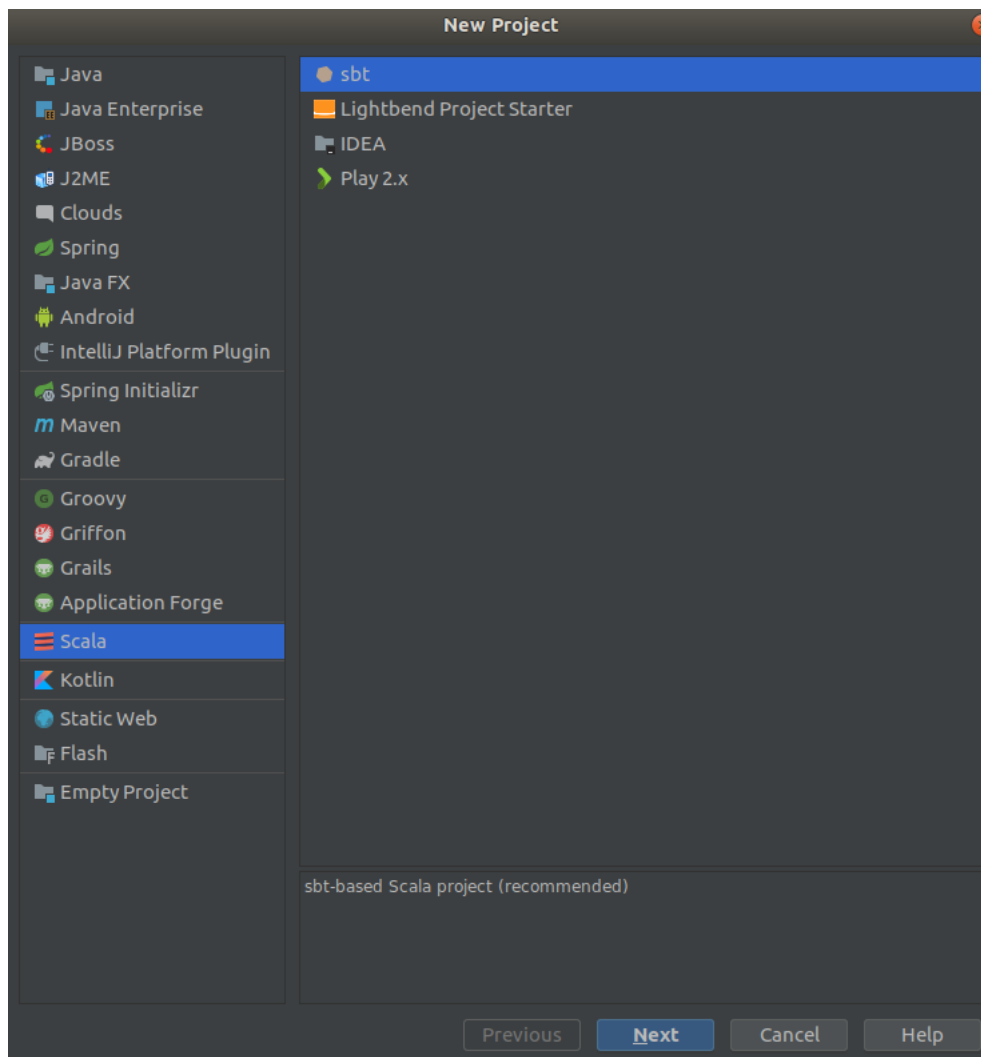
28 Setelah kebutuhan telah terpenuhi maka percobaan dapat dimulai. Berikut adalah langkah-langkah  
29 percobaan:

1. Pertama, buka IntelliJ dan buat sebuah project SBT seperti pada Gambar 3.3.



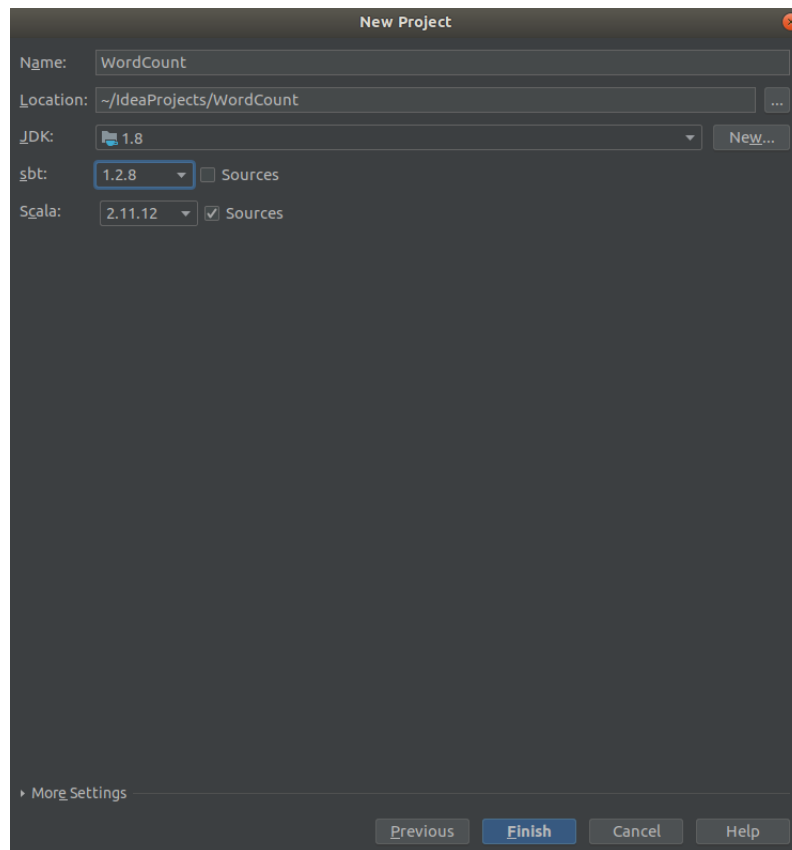
Gambar 3.3: ItelliJ IDEA

2. Setelah itu, pilih proyek Scala yang menggunakan sbt. Tekan tombol *next* seperti pada Gambar 3.4.



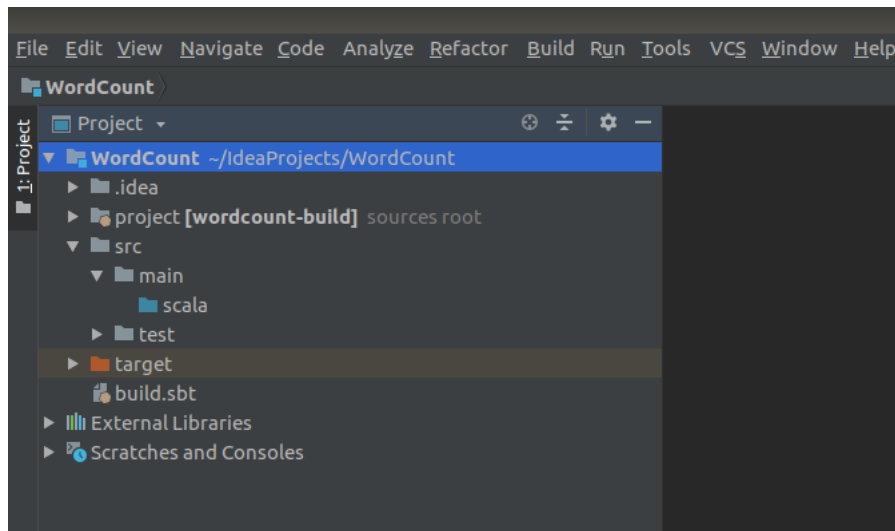
Gambar 3.4: Proyek sbt

- 1 Kemudian, beri nama proyek dengan nama WordCount dan pilih versi Sbt, Java, dan Scala yang sesuai
- 2 seperti pada Gambar 3.5.



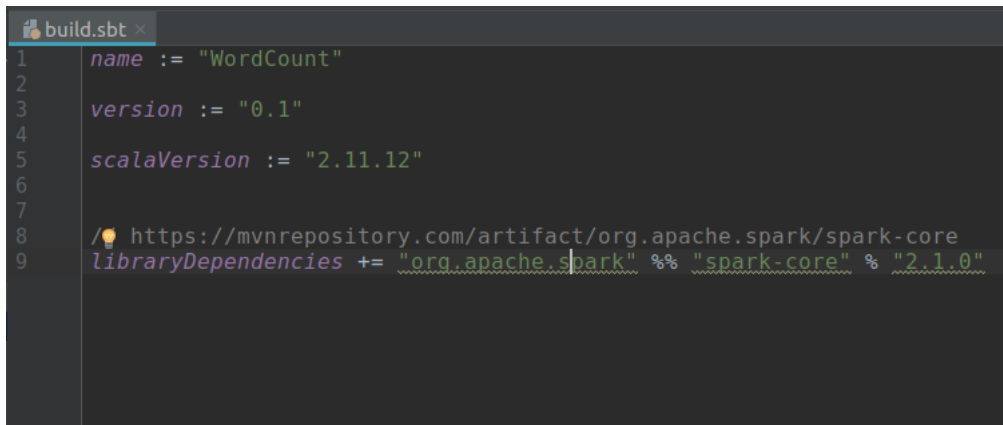
Gambar 3.5: Konfigurasi proyek

- 1 Hasil dari pembuatan proyek baru pada IntelliJ akan terlihat seperti pada Gambar 3.6.



Gambar 3.6: Struktur proyek

- 2 Setelah membuat proyek baru, buka file build.sbt dan tambahkan baris seperti pada Gambar 3.7.



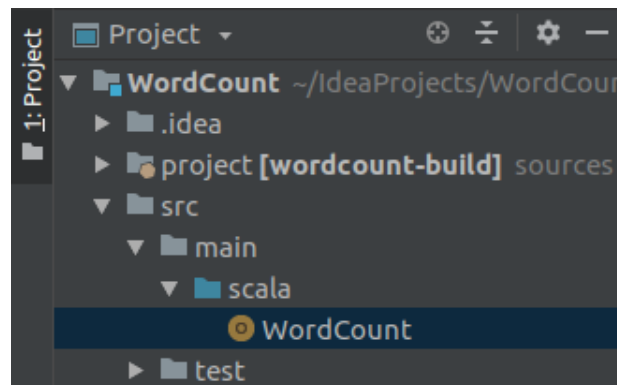
```

1  name := "WordCount"
2
3  version := "0.1"
4
5  scalaVersion := "2.11.12"
6
7
8  // https://mvnrepository.com/artifact/org.apache.spark/spark-core
9  libraryDependencies += "org.apache.spark" %% "spark-core" % "2.1.0"

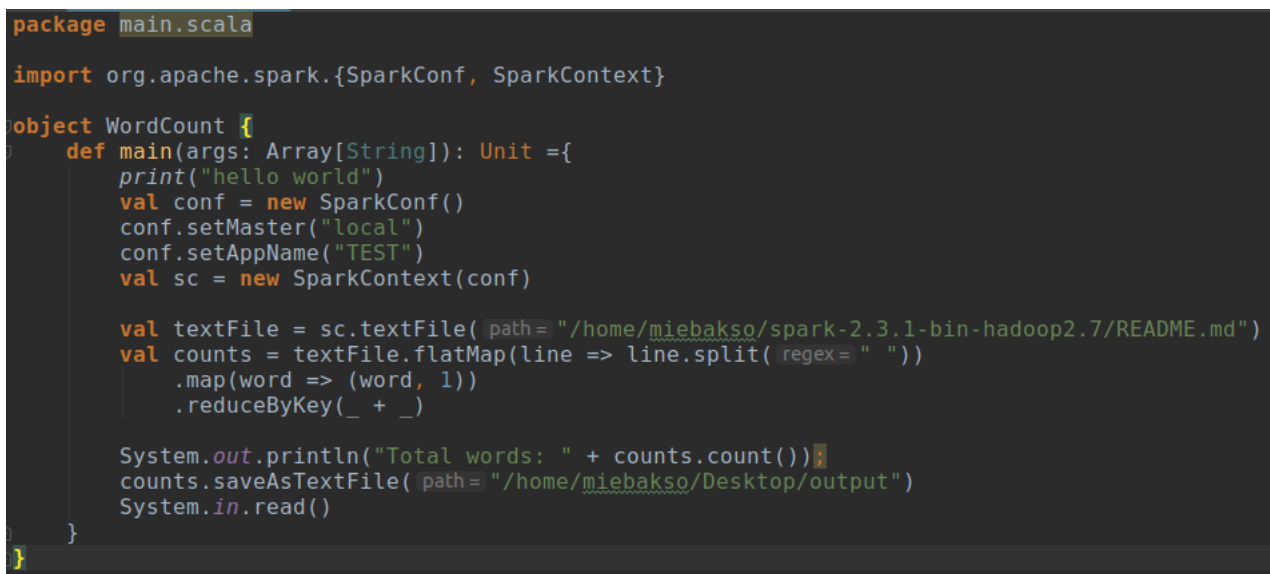
```

Gambar 3.7: Konfigurasi sbt

- 1 3. Tambahkan *object* WordCount pada proyek seperti pada Gambar3.8.

Gambar 3.8: *object* WordCount

- 2 Setelah itu, tambahkan kode berikut seperti pada Gambar3.9.



```

package main.scala

import org.apache.spark.{SparkConf, SparkContext}

object WordCount {
  def main(args: Array[String]): Unit = {
    print("hello world")
    val conf = new SparkConf()
    conf.setMaster("local")
    conf.setAppName("TEST")
    val sc = new SparkContext(conf)

    val textFile = sc.textFile(path = "/home/miebakso/spark-2.3.1-bin-hadoop2.7/README.md")
    val counts = textFile.flatMap(line => line.split(regex = " "))
      .map(word => (word, 1))
      .reduceByKey(_ + _)

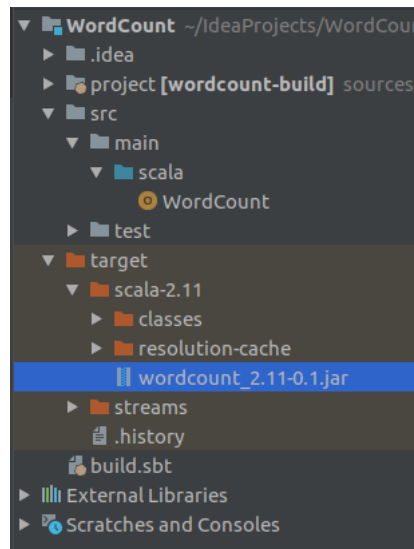
    System.out.println("Total words: " + counts.count());
    counts.saveAsTextFile(path = "/home/miebakso/Desktop/output")
    System.in.read()
  }
}

```

Gambar 3.9: Kode WordCount



- 1 4. Jalankan perintah 'sbt package' untuk meng-*compile* kode menjadi *executable* JAR seperti pada  
 2 Gambar 3.10. Hasil keluaran dapat dilihat pada Gambar 3.11.



Gambar 3.10: JAR

```
miebakso@black:~/IdeaProjects/WordCount$ sbt package
[info] Loading settings for project global-plugins from idea.sbt ...
[info] Loading global plugins from /home/miebakso/.sbt/1.0/plugins
[info] Loading project definition from /home/miebakso/IdeaProjects/WordCount/project
[info] Loading settings for project wordcount from build.sbt ...
[info] Set current project to WordCount (in build file:/home/miebakso/IdeaProjects/WordCount/)
[info] Compiling 1 Scala source to /home/miebakso/IdeaProjects/WordCount/target/scala-2.11/classes ...
[info] Done compiling.
[info] Packaging /home/miebakso/IdeaProjects/WordCount/target/scala-2.11/wordcount_2.11-0.1.jar ...
[info] Done packaging.
[success] Total time: 3 s, completed Apr 17, 2019 4:26:30 PM
miebakso@black:~/IdeaProjects/WordCount$
```

Gambar 3.11: Hasil perintah 'sbt package'

- 3 5. Setelah berhasil membuat JAR, masukan file JAR kepada *spark-submit* seperti pada Gambar 3.12.  
 4 Berikut adalah perintah yang harus dijalankan:

```
5 $ cd $SPARK_HOME
6 $ ./bin/spark-submit --class main.scala.WordCount --master local[1] \
7 /home/miebakso/IdeaProjects/WordCount/target/scala-2.11/wordcount_2.11-0.1.jar
```

```
miebakso@black:~/spark-2.3.1-bin-hadoop2.7$ ./bin/spark-submit --class main.scala.WordCount
--master local[2] /home/miebakso/IdeaProjects/WordCount/target/scala-2.11/wordcount_2.11-0.1.jar
2019-04-17 16:47:50 WARN Utils:66 - Your hostname, black resolves to a loopback address:
127.0.1.1; using 192.168.177.101 instead (on interface wlp5s0)
2019-04-17 16:47:50 WARN Utils:66 - Set SPARK_LOCAL_IP if you need to bind to another address
```

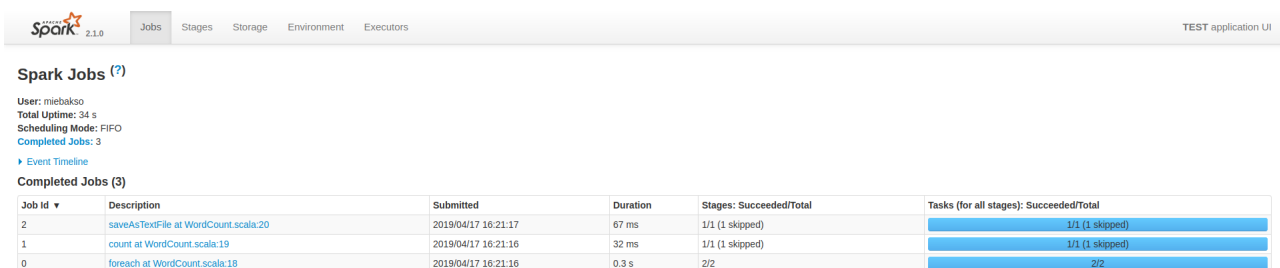
Gambar 3.12: Penggumpulan JAR kepada *spark-submit*

- 1 Hasil tahap-tahap proses dari program dapat dilihat pada Spark UI dengan membuka alamat yang  
 2 digaris bawah biru pada Gambar 3.13

```
SparkEnv: Registering MapOutputTracker
SparkEnv: Registering BlockManagerMaster
BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper
BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
DiskBlockManager: Created local directory at /tmp/blockmgr-4c6caad2-7b7d-42ad-
MemoryStore: MemoryStore started with capacity 1951.2 MB
SparkEnv: Registering OutputCommitCoordinator
Utils: Successfully started service 'SparkUI' on port 4040.
SparkUI: Bound SparkUI to 0.0.0.0, and started at http://192.168.177.101:4040
Executor: Starting executor ID driver on host localhost
Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlock
NettyBlockTransferService: Server created on 192.168.177.101:41353
BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy for
BlockManagerMaster: Registering BlockManager BlockManagerId(driver, 192.168.17
BlockManagerMasterEndpoint: Registering block manager 192.168.177.101:41353 wi
```

Gambar 3.13: Alamat Spark UI

- 3 Spark UI menggambarkan tahap-tahap proses program. Tampilan dari Spark UI dapat dilihat pada  
 4 Gambar 3.14.



The screenshot shows the Spark UI interface with the 'Jobs' tab selected. It displays details for a job named 'TEST application UI'. The interface includes a header with the Spark logo and version (2.1.0), and navigation tabs for Jobs, Stages, Storage, Environment, and Executors. Below the header, it shows 'Spark Jobs (?)' with user information (User: njebakso), total uptime (34 s), scheduling mode (FIFO), and completed jobs (3). A table titled 'Completed Jobs (3)' lists the job details.

Job id	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
2	saveAsTextFile at WordCount.scala:20	2019/04/17 16:21:17	67 ms	1/1 (1 skipped)	1/1 (1 skipped)
1	count at WordCount.scala:19	2019/04/17 16:21:16	32 ms	1/1 (1 skipped)	1/1 (1 skipped)
0	foreach at WordCount.scala:18	2019/04/17 16:21:16	0.3 s	2/2	2/2

Gambar 3.14: Spark UI

## BAB 4

### ANALISIS DAN PERANCANGAN

Pada bab ini, akan dijelaskan hal-hal yang dilakukan dalam pengembangan *Agglomerative Hierarchical Clustering* untuk Spark. Pengembangan dilakukan untuk mencapai tujuan yaitu mendapatkan pola dari dataset yang diolah. Pola yang ingin didapatkan meliputi perhitungan rata-rata, nilai maksimum, nilai minimum dan nilai standar deviasi dari setiap atribut yang ada pada data. Selain itu, perlu didapatkan juga jumlah anggota pada setiap *cluster* yang dihasilkan dari algoritma *Hierarchical Agglomerative Clustering*.

#### 4.1 Analisis Masalah

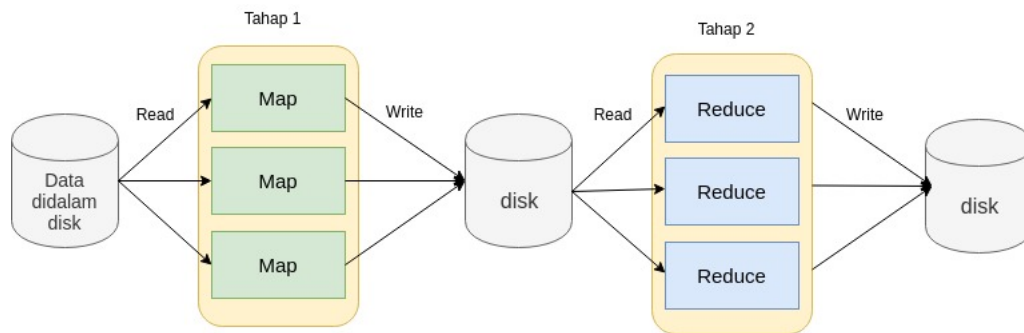
Bagian ini menjelaskan masalah dari penelitian ini, analisis algoritma *Hierarchical Agglomerative Clustering* dan analisis masukan.

##### 4.1.1 Identifikasi Masalah

Dalam bidang *big data*, volume data yang sangat besar harus disimpan dalam tempat penyimpanan yang sangat besar. Volume data *big data* dapat mencapai *peta bytes*. Volume yang terlalu besar akan meningkatkan biaya dan menghabiskan tempat penyimpanan data. Volume data perlu direduksi agar menghemat tempat dan biaya.

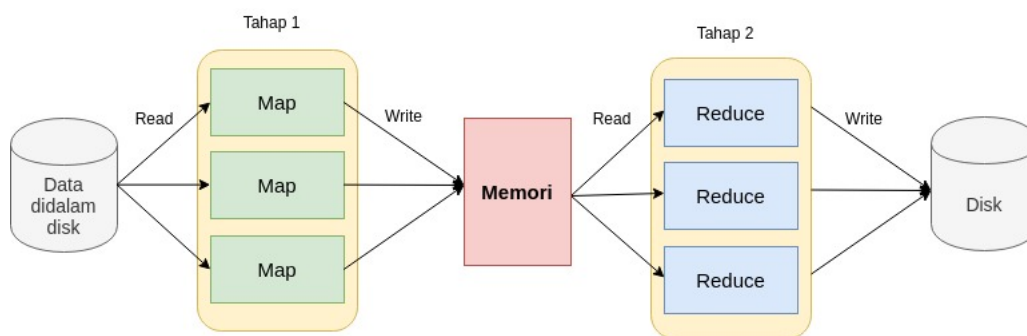
Hadoop MapReduce dan algoritma *Agglomerative Hierarchical Clustering* dapat digabungkan sebagai solusi untuk mereduksi data. Algoritma *Agglomerative* dapat mereduksi data dengan mengambil pola-pola dari *clusters* yang dibentuk. Sistem terdistribusi Hadoop membantu dalam proses membagikan dan memecah tugas agar dapat dikerjakan secara paralel. Dengan begitu, proses reduksi data dengan algoritma *Agglomerative* menjadi lebih cepat.

Tetapi Hadoop masih terlalu lambat dalam mereduksi data. Hal ini disebabkan karena Hadoop banyak melakukan penulisan dan pembacaan kepada disk. Proses *disk I/O* pada Hadoop sangat tinggi dan menyebabkan algoritma *Agglomerative* berjalan sangat lambat pada Hadoop. Pada setiap tahap, Hadoop akan menuliskan hasilnya kepada *disk* dan akan dibaca kembali oleh tahap selanjutnya dari *disk* seperti pada Gambar 4.1.



Gambar 4.1: Penulisan kepada disk di MapReduce

- 1 Solusinya adalah menggabungkan sistem terdistribusi lainnya dengan algoritma *Agglomerative* untuk
- 2 mereduksi data. Spark yang dapat menyimpan data pada memori dapat menggantikan Hadoop MapReduce.
- 3 Kecepatan memori lebih cepat dibanding *disk* merupakan salah satu faktor mengapa Spark akan mempro-
- 4 ses data dengan kecepatan yang lebih tinggi. Pembacaan dan penulisan akan dilakukan kepada memori.
- 5 Gambar 4.2 adalah contoh ilustrasi tahap proses data di Spark.



Gambar 4.2: Penulisan kepada memori di Spark

#### 6 4.1.2 Analisis *Hierarchical Agglomerative Clustering* MapReduce

- 7 Sebelum melakukan perancangan, penulis terlebih dahulu mempelajari algoritma *Hierarchical Agglomerative*
- 8 *Clustering* pada Hadoop. Algoritma *Hierarchical Agglomerative Clustering* pada MapReduce dibagi menjadi
- 9 dua bagian. Bagian pertama terkait tahap *map* dan bagian kedua terkait tahap *reduce*. Tahap *map* bertujuan
- 10 untuk membagi rata data menjadi beberapa partisi agar setiap *reducer* mendapatkan pekerjaan yang hampir
- 11 rata dengan *reducer* yang lainnya. Tahap *map* akan dijelaskan pada *pseudocode* berikut ini 1:

**Algorithm 1:** Algoritma *Mapper***Masukan :** Data mentah (**TO**), jumlah partisi ( $n$ )**Keluaran :**  $key$  = sebuah bilangan bulat  $\in \{1 \dots n\}$ ,  $value$  = teks dari sekumpulan nilai atribut yang telah diproses sebelumnya**Deskripsi :** memecah **TO** dengan memberi bilangan acak untuk setiap objek**1 begin****2     value**  $\leftarrow$  membaca baris dan memproses atributnya**3     key**  $\leftarrow$  sebuah bilangan acak  $k$ , dimana  $1 \leq k \leq n$ **4     mengembalikan** pasangan  $\langle key, value \rangle$  sebagai hasil**5 end**

Tahap *reduce* bertujuan untuk mereduksi data. Pada tahap ini dendrogram akan dibangun dari hasil tahap *map*. Setelah membangun *dendrogram*, *dendrogram* akan dipotong untuk menghasilkan *clusters*. Kemudian, pola akan dihitung dari *clusters* dan disimpan kepada file. Tahap *reduce* akan dijelaskan dengan *pseudocode* berikut ini 2:

**Algorithm 2:** Algoritma *reducer*

**Masukan :** pasangan  $\langle key, value \rangle$  dari mapper dimana semua *value*-nya memiliki nilai *key* yang sama, *maxObject*, *distType*  $\in \{single, complete, means\}$ , *cut-off distance*  $\{co\}$

**Keluaran :** pola *cluster*, *c*

**Deskripsi :** Membuat *dendrogram* dari hasil *map* sesuai dengan batasan yang diberikan, membatasi jumlah objek yang akan diolah menjadi *dendrogram* berdasarkan *maxObject*, menghitung pola dari *cluster* berdasarkan nilai *co*, menuliskan hasil pola kepada file

```

1  begin
2      listTrees  $\leftarrow []$ 
3      foreach pasangan  $\langle key, value \rangle$  do
4          node  $\leftarrow value$ 
5          tambahkan node kepada listTrees
6          isProcessed  $\leftarrow false$ 
7          if listTrees.length == maxObject then
8              bangun dendrogram dari listTrees berdasarkan tipe distType
9              bentuk clusters dari dendrogram berdasarkan nilai co
10             hitung pola c dari setiap cluster yang dibentuk dan simpan hasil kepada file
11             kosongkan listTrees
12             isProcessed  $\leftarrow True$ 
13         end
14     end
15     if isProcessed == false then
16         bangun dendrogram dari listTrees berdasarkan tipe distType
17         bentuk clusters dari dendrogram berdasarkan nilai co
18         hitung pola c dari setiap cluster yang dibentuk dan simpan hasil kepada file
19     end
20 end

```

### 4.1.3 Analisis Masukan dan Keluaran

Dalam melakukan perancangan perlu diketahui terlebih dahulu kebutuhan perangkat lunak. Perangkat lunak yang dirancang harus dapat menangani masukan yang diberikan seperti contoh di bawah. Setiap baris mewakili sebuah objek beserta atributnya. Atribut dipisahkan dengan tanda koma. Setiap atribut merupakan bilangan desimal. Setiap objek dapat memiliki lebih dari satu atribut.

```

97.92268076905681,95.67804892782392
15.875897725375477,81.36427207827654
15.825886365695096,6.163384415958262
69.28295038155534,85.36655250595662
10.032110782002924,98.13534474918522
38.53402755308164,96.99987611939603
45.17834148867077,5.96338806209017

```

1 91.66074344459808, 15.182927773314525

2 . . . .

3 . . . .

4 Selain itu, perangkat lunak harus dapat menghasilkan pola seperti berikut:

5 1. Jumlah objek pada *cluster*.

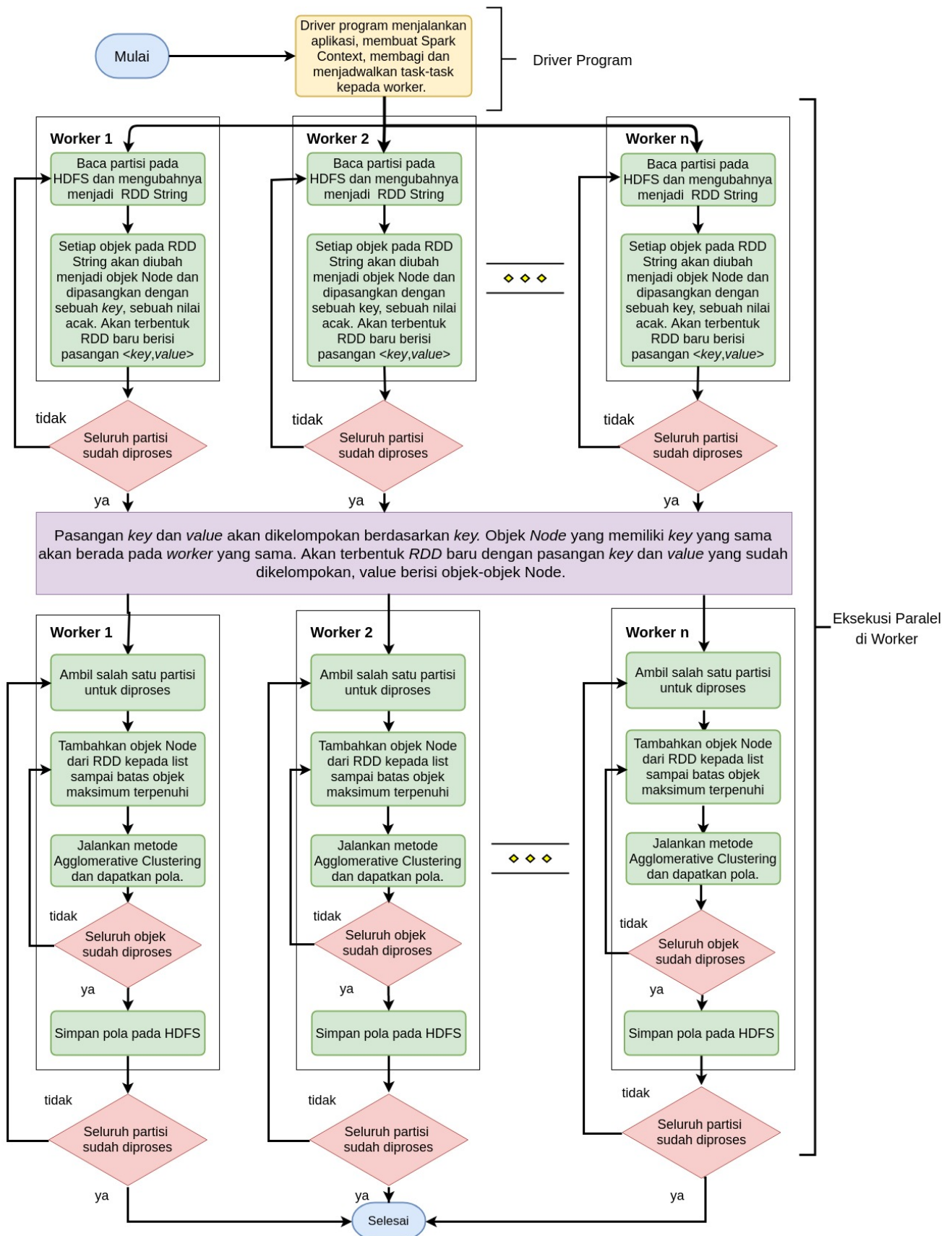
6 2. Nilai minimum setiap atribut pada *cluster*.

7 3. Nilai maksimum setiap atribut pada *cluster*.

8 4. Nilai rata-rata setiap atribut pada *cluster*.

9 5. Nilai standar deviasi setiap atribut pada *cluster*.

#### 1 4.1.4 Diagram Alur

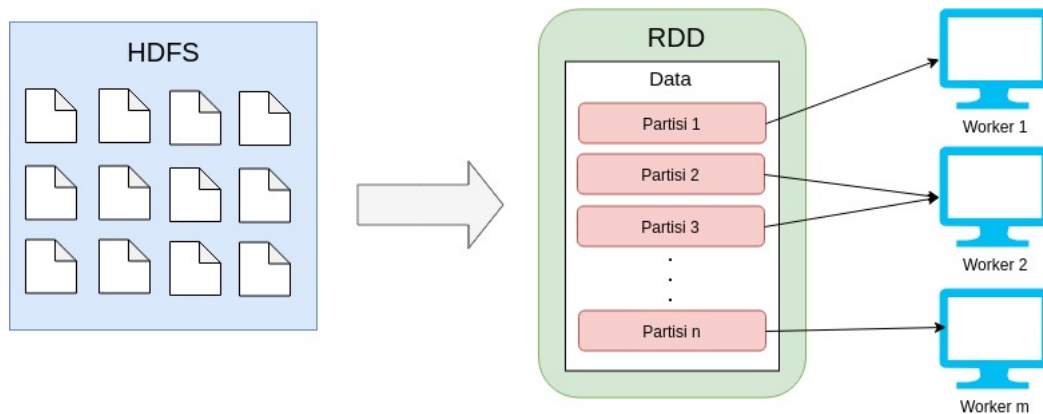


Gambar 4.3: Diagram alur perangkat lunak



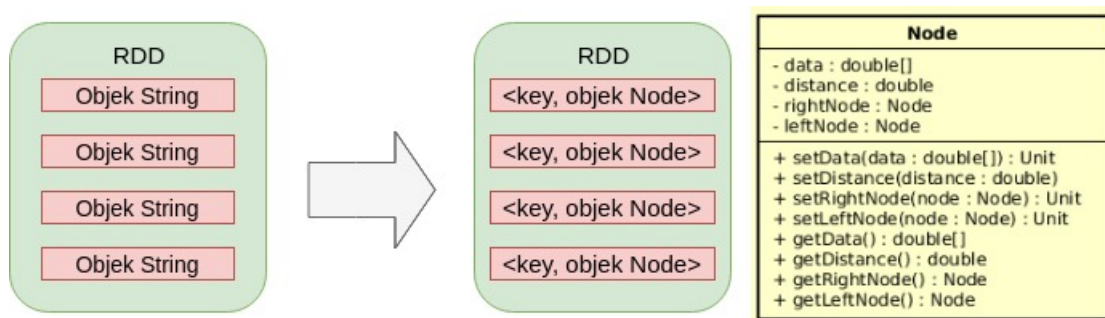
1 Diagram alur pada Gambar 4.3 digunakan untuk menjelaskan alur perangkat lunak. Berikut adalah penjelasan  
 2 alur perangkat lunak:

- 3 1. Pertama-tama aplikasi akan dijalankan pada *driver program*. Kemudian *Spark Context* akan dibuat  
 4 dan operasi-operasi pada aplikasi diubah menjadi *task-task*. *Task-task* tersebut akan dibagikan dan  
 5 dijadwalkan kepada *worker* oleh *driver program*.
- 6 2. Kemudian, *worker* akan membaca partisi HDFS yang ditentukan oleh *driver program*. *Worker* akan  
 7 membaca *blocks* tersebut sebagai RDD bertipe String seperti pada Gambar 4.4.

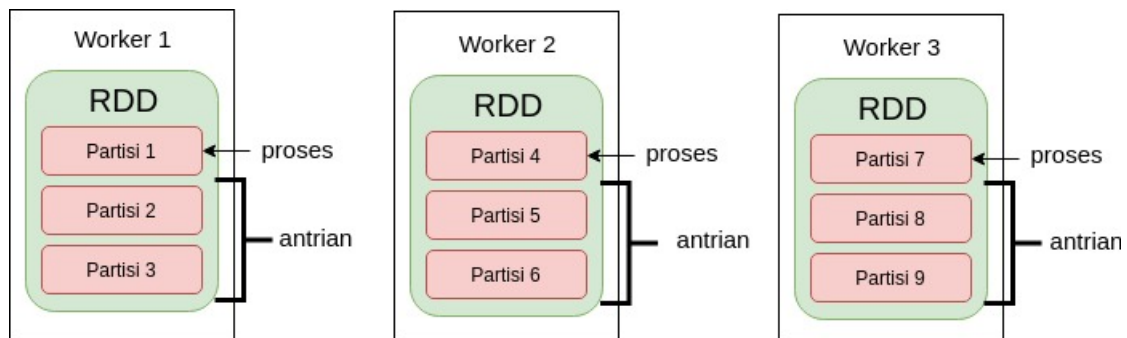


Gambar 4.4: Partisi RDD

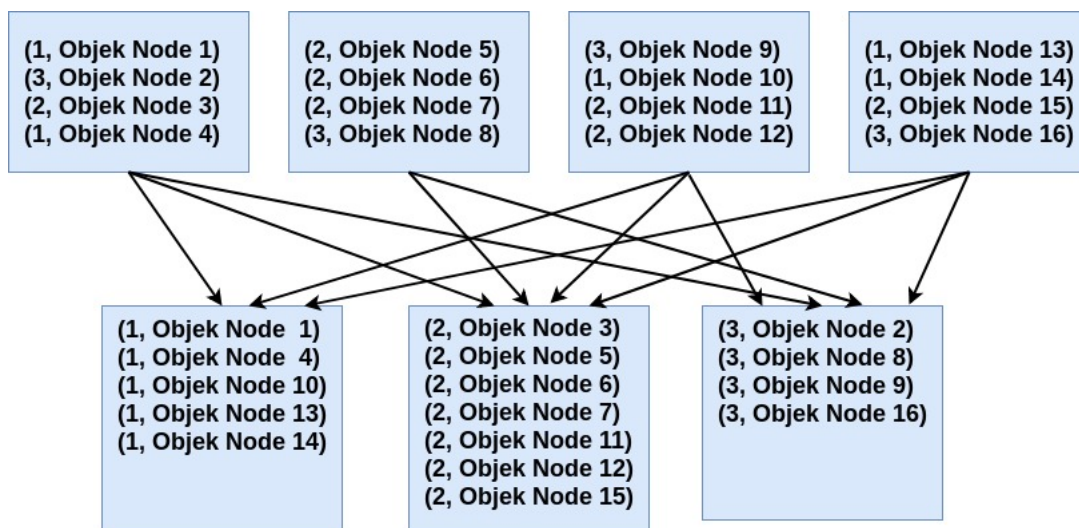
- 8 3. Selanjutnya, setiap objek pada RDD bertipe *String* akan diolah menjadi objek *Node*. Objek *Node*  
 9 akan dipasangkan dengan sebuah *key*. *Key* merupakan bilangan acak antara 1 sampai *n*. Bilangan *n*  
 10 adalah jumlah partisi yang ditentukan oleh pengguna. Akan dihasilkan RDD baru berisi pasangan  
 11  $\langle \text{key}, \text{value} \rangle$  seperti pada Gambar 4.5. *Worker* akan memproses satu partisi dan melanjutkannya ketika  
 12 selesai dengan partisi yang sedang diproses seperti pada Gambar 4.6.



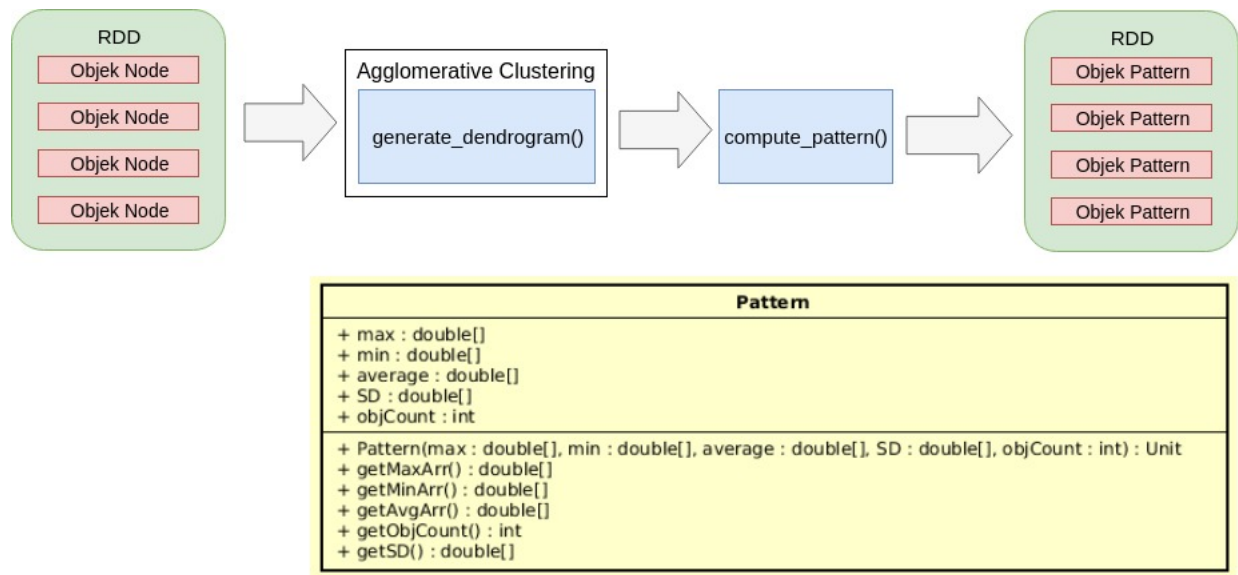
Gambar 4.5: RDD parsing dan kelas Node

Gambar 4.6: *Worker* memproses partisi

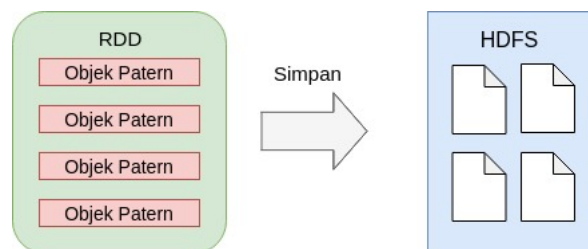
4. Setelah itu, akan terjadi pengelompokkan berdasarkan *key* yang sama. Akan terjadi perpindahan data dari satu *worker* kepada *worker* lainnya. Objek *Node* dengan *key* yang sama akan berada pada memori *worker* yang sama seperti pada Gambar 4.7.

Gambar 4.7: Pengelompokkan *Node* berdasarkan *key*

5. Setelah data dikelompokkan berdasarkan *key*, objek *Node* akan dimasukkan kepada sebuah *list* sampai batas objek maksimum yang ditentukan oleh pengguna. Selanjutnya metode *agglomerative clustering* akan dipanggil seperti pada Gambar 4.8. Metode ini akan membangun sebuah *dendrogram* menggunakan algoritma HAC. *Dendrogram* akan dipotong untuk menghasilkan *cluster-cluster*. Setiap *cluster* akan dicari polanya. Pola akan dikembalikan sebagai hasilnya. Langkah ini akan diulangi sampai seluruh objek pada partisi telah diproses.

Gambar 4.8: Proses reduksi dan kelas *Pattern*

- 1 6. Terakhir, pola akan disimpan pada HDFS seperti pada Gambar 4.9. Bila semua partisi sudah diproses,
- 2 perangkat lunak akan berhenti. Bila masih ada partisi yang tersisa, ulangi langkah sebelumnya sampai
- 3 seluruh partisi telah diproses.



Gambar 4.9: Penyimpanan pola pada HDFS

#### 4 4.1.5 Analisis *Hierarchical Agglomerative Clustering* pada Spark

- 5 Setelah mempelajari algoritma *Hierarchical Agglomerative Clustering* pada MapReduce, *format* masukan
- 6 yang harus diproses dan keluaran yang harus dihasilkan, berikut adalah penjelasan *pseudocode* algoritma
- 7 *map* dan *reduce* untuk Spark:

**Algorithm 3:** Algoritma *Map*

**Masukan :** dataset (*A*) bertipe RDD[String], jumlah partisi (*n*)

**Keluaran :** *DN* = dataset baru bertipe RDD[<key,Node>]

**Deskripsi :** Melakukan *parsing* dan memasangkan key untuk setiap elemen pada RDD *A*.

Mengembalikan RDD baru bertipe <key,Node>

1 **begin**

2     *DN*  $\leftarrow$  RDD bertipe <key,Node> yang kosong

3     **foreach** *line* pada *A* **do**

4         *node*  $\leftarrow$  node baru

5         *split*  $\leftarrow$  split *line* berdasarkan delimiter "," dan konversi menjadi double

6         *node.setData(split)*

7         *randomKey*  $\leftarrow$  hasilkan bilangan acak antara 1 sampai dengan *n*

8         *DN*  $\leftarrow$  *DN* join <*randomKey,node*>

9     **end**

10    return *DN*

11 **end**

**Algorithm 4:** Algoritma *Reduce*

**Masukan :** (*DN*) RDD[<key,Node>] hasil dari mapper, jumlah objek maksimum (*MX*), tipe metode yang dipakai (*distType*)  $\in \{single, complete, centroid\}$ , dan *cut-off distance* (*co*)

**Deskripsi :** Membuat *dendrogram* dari hasil *map* sesuai dengan batasan yang diberikan, membatasi jumlah objek yang akan diolah menjadi *dendrogram* berdasarkan *MX*, memotong *dendrogram* berdasarkan nilai *co*, mendapatkan pola *pt* dari potongan *cluster*, menyimpan pola-pola pada HDFS

```

1 begin
2   broadcast nilai MX, distType, dan co
3   objectList  $\leftarrow$  [] array kosong bertipe Node
4   patterns  $\leftarrow$  RDD bertipe Pattern untuk mengumpulkan pola hasil reduksi
5   foreach elemen in DN.value do
6     objectList  $\leftarrow$  objectList join elemen
7     isProcessed  $\leftarrow$  false
8     if count(objectList) == MX then
9       dendrogram  $\leftarrow$  generate_dendrogram(objectList, distType)
10      pt  $\leftarrow$  compute_pattern(dendrogram, co)
11      patterns  $\leftarrow$  pattern join pt
12      isProcessed  $\leftarrow$  true
13      kosongkan objectList
14    end
15  end
16  if isProcessed == false then
17    dendrogram  $\leftarrow$  generate_dendrogram(objectList, distType)
18    pt  $\leftarrow$  compute_pattern(dendrogram, co)
19    patterns  $\leftarrow$  pattern join pt
20  end
21  foreach pattern in patterns do
22    simpan pattern pada HDFS
23  end
24 end

```

**Function** generate\_dendrogram(*objectList*, *distType*):

**Masukan :** list objek-objek *objectList*, tipe metode *distType*

**Keluaran :** dendrogram

**Deskripsi :** Membangun *dendrogram* dari list objek sesuai dengan nilai *distType* yang diberikan

```

1  begin
2      distanceMatrix  $\leftarrow$  [][] array double untuk merepresentasikan jarak antara cluster
3      nodeListCluster  $\leftarrow$  [] array bertipe List<Node> merepresentasikan cluster
4      dendrogram  $\leftarrow$  [] array bertipe node untuk merepresentasikan dendrogram cluster
5      i  $\leftarrow$  0
6      foreach node in objectList do
7          nodeListCluster[i]  $\leftarrow$  nodeListCluster[i] join node
8          dendrogram  $\leftarrow$  dendrogram join node
9          i  $\leftarrow$  i + 1
10     end
11     i  $\leftarrow$  1
12     j  $\leftarrow$  0
13     for i < distanceMatrix.length do
14         for j < i do
15             distanceMatrix[i][j]  $\leftarrow$  findMinDist(nodeListCluster[i],nodeListCluster[j],distType)
16             j  $\leftarrow$  j + 1
17         end
18         i  $\leftarrow$  i + 1
19     end
20     while dendrogram.length != 1 do
21         i  $\leftarrow$  1
22         j  $\leftarrow$  0
23         x  $\leftarrow$  0
24         y  $\leftarrow$  0
25         temp  $\leftarrow$  0
26         result  $\leftarrow$  Double.MaxValue
27         for i < distanceMatrix.length do
28             for j < i do
29                 temp  $\leftarrow$  distanceMatrix[i][j]
30                 if temp < result then
31                     result  $\leftarrow$  temp
32                     x  $\leftarrow$  i
33                     y  $\leftarrow$  j
34                 end
35                 j  $\leftarrow$  j + 1
36             end
37             i  $\leftarrow$  i + 1
38         end
39         nodeListCluster[y]  $\leftarrow$  nodeListCluster[y] join nodeListCluster[x]
40         nodeListCluster.remove(x)
41         newNode  $\leftarrow$  merupakan Node baru
42         newNode.setDistance(distanceMatrix[x][y])
43         newNode.setLeftNode(dendrogram[y])
44         newNode.setRightNode(dendrogram[x])
45         dendrogram[y]  $\leftarrow$  newNode
46         dendrogram.remove(x)
47         recalculateMatrix(distanceMatrix, nodeListCluster,x,y)
48     end
49     return dendrogram[0]
50 end

```

**Function** findMinDist(*listA*, *listB*, *distType*):

**Masukan :** list objek Node *listA*, list objek Node *listB*, tipe metode *distType*

**Keluaran :** nilai double

**Deskripsi :** Mencari jarak antara cluster A dan B berdasarkan tipe jarak yang digunakan

```

1  begin
2      if distType adalah Single Linkage then
3          min  $\leftarrow$  Double.MaxValue
4          result  $\leftarrow$  0
5          foreach nodeA in listA do
6              foreach nodeB in listB do
7                  result  $\leftarrow$  Cari jarak euclidean antara nodeA dan nodeB
8                  if result < min then
9                      min  $\leftarrow$  result
10                 end
11             end
12         end
13     else if distType adalah Complete Linkage then
14         max  $\leftarrow$  Double.MinValue
15         result  $\leftarrow$  0
16         foreach nodeA in listA do
17             foreach nodeB in listB do
18                 result  $\leftarrow$  Cari jarak euclidean antara nodeA dan nodeB dengan eu
19                 if result < max then
20                     max  $\leftarrow$  result
21                 end
22             end
23         end
24     else
25         centroidA  $\leftarrow$  cari centroid dari listA
26         centroidB  $\leftarrow$  cari centroid dari listB
27         result  $\leftarrow$  Cari jarak euclidean antara centroidA dan centroidB
28     end
29     return result
30 end

```

**Function** recalculateMatrix(*distanceMatrix*, *nodeListCluster*, *x*, *y*):

**Masukan** : jarak antara cluster *distanceMatrix*, array berisi lisT Node *nodeListCluster*, index *x*, index *y*

**Deskripsi** : Melakukan kalkulasi ulang natara clusters dan cluster baru

```
1  begin
2      distanceMatrix.remove(x)
3      i ← i + 1
4      for i < distanceMatrix.length do
5          distanceMatrix[i].remove(x)
6      end
7      i ← y + 1
8      for i < distanceMatrix.length do
9          distanceMatrix ← findMinDist(nodeListCluster[i], nodeListCluster[y])
10     end
11 end
```



**Function** `compute_pattern(dendrogram, co):`

**Masukan :** *dendrogram*, *cut-off distance co*

**Keluaran :** pola-pola dari seluruh potongan cluster

**Deskripsi :** Memotong *dendrogram* menjadi beberapa *clusters* berdasarkan nilai *co*, mendapatkan pola dari setiap *cluster*

```

1  begin
2      bfs  $\leftarrow$  [] array kosong bertipe Node
3      clusters  $\leftarrow$  [] array kosong untuk menyimpan hasil potongan dari dendrogram
4      bfs.add(dendrogram)
5      dist  $\leftarrow$  co * dendrogram.distance
6      while bfs tidak kosong do
7          node  $\leftarrow$  bfs.remove(0)
8          if node.distance  $\leq$  dist then
9              clusters.add(node)
10         else
11             left  $\leftarrow$  node.left
12             right  $\leftarrow$  node.right
13             if left  $\neq$  null then
14                 bfs.add(left)
15             end
16             if right  $\neq$  null then
17                 bfs.add(right)
18             end
19         end
20     end
21     patterns[]
22     foreach cluster in clusters do
23         p  $\leftarrow$  dapatkan pola dari setiap cluster
24         patterns.add(p)
25     end
26     return patterns
27 end

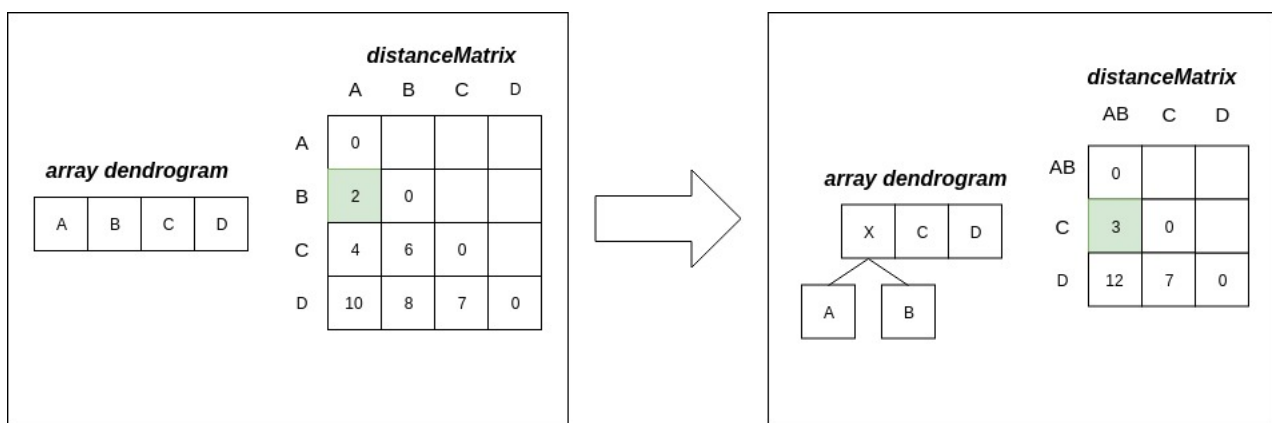
```

Algoritma Map 3 ini bertujuan untuk melakukan *parsing* terhadap masukan yang diberikan. Masukan yang diberikan berupa RDD[String]. Setiap elemen pada RDD[String] akan di-*parsing* menjadi objek *Node* dan dipasangkan dengan *key*, yaitu bilangan acak antara 1 sampai *n*. Bilangan *n* merupakan jumlah partisi yang diberikan oleh pengguna. Pertama, elemen pada RDD[String] yang berupa *String* akan dipecah berdasarkan *delimiter* ",", dan dikonversi menjadi bilangan pecahan. Hasilnya merupakan *array* bertipe *double* yang menjadi atribut objek *Node*. Kemudian akan diberikan bilangan acak antara 1 sampai *n*. Bilangan tersebut akan dipasangkan kepada objek *Node*. Pasangan *<key,Node>* kemudian akan tambahkan kepada RDD[*<key,Node>*]. RDD[*<key,Node>*] dikembalikan sebagai hasil dan menjadi masukan untuk tahap *reduce*.

Algoritma 4 mengenai *reduce* bertujuan untuk membangun *dendrogram* dan mengembalikan pola-pola

bertipe RDD[Pattern] sebagai hasilnya. Pertama-tama nilai  $MX$ ,  $distType$ ,  $co$  akan di-*broadcast* agar setiap *worker* memiliki nilai tersebut. *Variable objectList* dibuat untuk menampung *Nodes* yang akan dibangun menjadi *dendrogram*. *Node* pada RDD[<key,Node>] akan ditambahkan kepada *objectList* sampai batas jumlah *Node* pada *objectList* sama dengan nilai  $MX$ . Kemudian, fungsi *generate\_dendrogram(objectList, distType)* akan dipanggil untuk membangun *dendrogram*. Hasil dari fungsi tersebut yaitu sebuah *dendrogram* yang dijadikan sebagai masukan untuk fungsi *compute\_pattern(dendrogram,co)*. Fungsi ini memotong *dendrogram* menjadi *cluster-cluster* dan mencari pola dari setiap *cluster*. Pola atau *Pattern* akan ditambahkan kepada *variable pattern* (RDD[Pattern]) yang akan dikembalikan sebagai hasil. *Variable isProcessed* pada baris 16 untuk memastikan bahwa setiap elemen pada *objectList* sudah diproses.

Fungsi *generate\_dendrogram* pada Algoritma 4 digunakan untuk membangun *dendrogram*. Fungsi ini akan menerima *objectList* sebagai masukan. Pertama-tama *array distanceMatrix* harus diinisialisasi dan dihitung jarak antara objeknya menggunakan  $distType$  yang ditentukan. Kemudian, *array dendrogram* diisi dengan objek-objek pada *objectList*. Untuk membangun *dendrogram*, gabungkan objek yang memiliki nilai terkecil pada *distanceMatrix* seperti pada Gambar 4.10. Setelah menggabungkan dua buah objek, objek pada *dendrogram* akan berkurang satu dan *distanceMatrix* harus dihitung ulang berdasarkan  $distType$ .



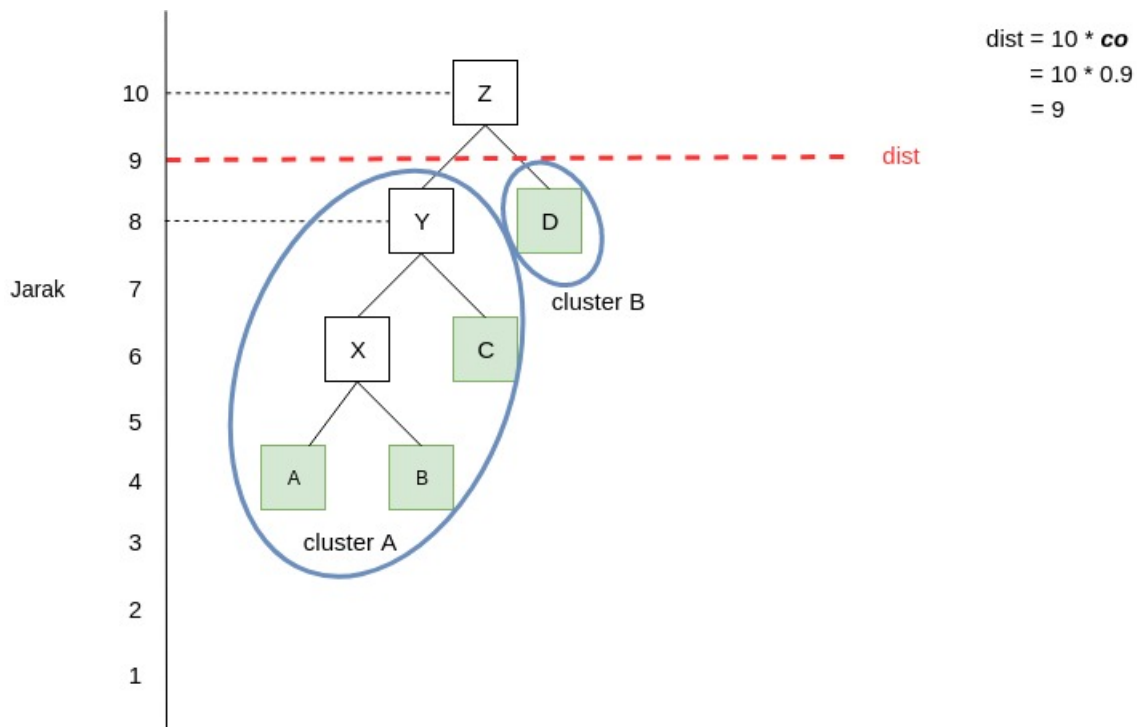
Gambar 4.10: Contoh perhitungan matriks dan pembentukan dendrogram

Fungsi *findMinDist* pada Algoritma 4 digunakan untuk mencari jarak antara *cluster* berdasarkan  $distType$  yang dipilih. Setiap anggota pada *cluster* akan dibandingkan dengan anggota di *cluster* lainnya. Berdasarkan  $distType$  maka akan dicari nilai minimum, maksimum, atau jarak antara *centroid*-nya.

Fungsi *recalculateMatrix* pada Algoritma 4 digunakan menghitung ulang jarak antara *cluster* baru dengan *cluster* lainnya. *Cluster* baru akan dihitung jaraknya berdasarkan  $distType$  yang dipilih. Bila  $distType$  yang dipilih adalah *Single Linkage* maka jarak minimum dari perbandingan anggota dari kedua *cluster* akan diambil sebagai hasilnya.

Fungsi *compute\_pattern* pada Algoritma 4 digunakan untuk mendapatkan pola dari *cluster*. Fungsi ini menerima hasil *dendrogram* dari fungsi *generate\_dendrogram*, beserta nilai *cut-off distance* sebagai masukkannya. Pertama-tama *dendrogram* yang diwakili dengan struktur *tree* akan ditelusuri di setiap tingkatnya. Jarak pada setiap tingkat akan di cek. Bila jarak sudah kurang dari jarak hasil perkalian  $co$  dengan tinggi *dendrogram*, maka *dendrogram* akan dipotong untuk menghasilkan potongan *clusters*. Setelah itu, pola dari

- 1 setiap cluster akan dicari. Pola didapatkan dengan mencari nilai minimum, maksimum, rata-rata dan standard
- 2 deviasi dari setiap attribute pada *cluster*. Pola akan dikembalikan sebagai hasil.

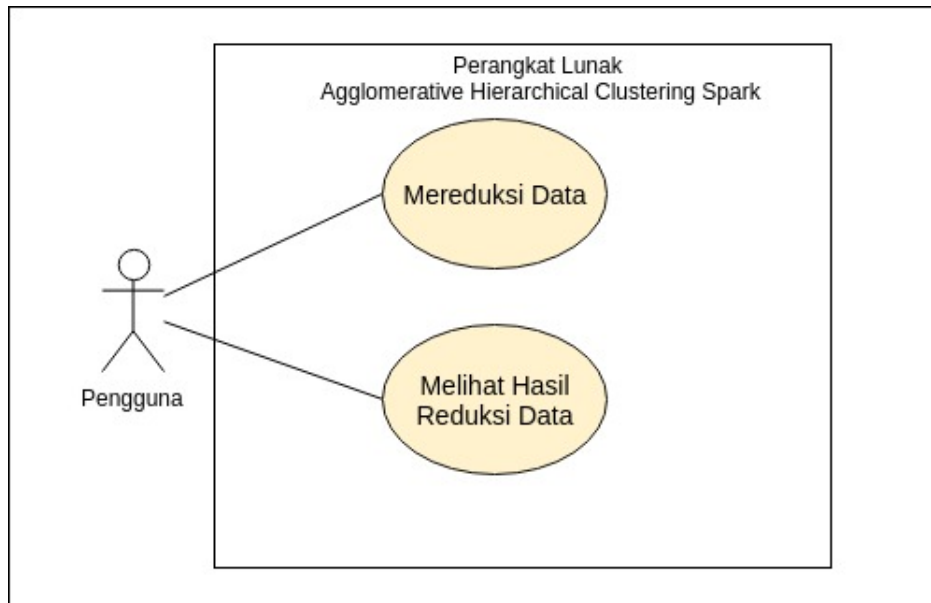
Gambar 4.11: Contoh pemotongan *dendrogram*

## 3 4.2 Perancangan Perangkat Lunak

- 4 Bagian ini menjelaskan perancangan perangkat lunak. Perancangan termasuk diagram *use case*, skenario,
- 5 diagram kelas, dan rancangan antarmuka.

### 6 4.2.1 Diagram *Use Case* dan Skenario

- 7 Diagram *use case* merupakan sebuah pemodelan untuk perilaku dari perangkat lunak yang akan dibuat.
- 8 Diagram *use case* digunakan untuk mengetahui fungsi apa saja yang ada dalam perangkat lunak. Fungsi-
- 9 fungsi dari perangkat lunak akan dioperasikan oleh satu pengguna. Cara kerja dan perilaku dari perangkat
- 10 lunak dijelaskan dalam bentuk diagram *use case* yang dapat dilihat pada Gambar 4.12.



Gambar 4.12: Diagram *use case* perangkat lunak *Hierarchical Agglomerative Clustering*

Berdasarkan gambar diagram *use case* di atas, berikut merupakan beberapa skenario yang dapat terbentuk:

1. Nama *use case*: Mereduksi data

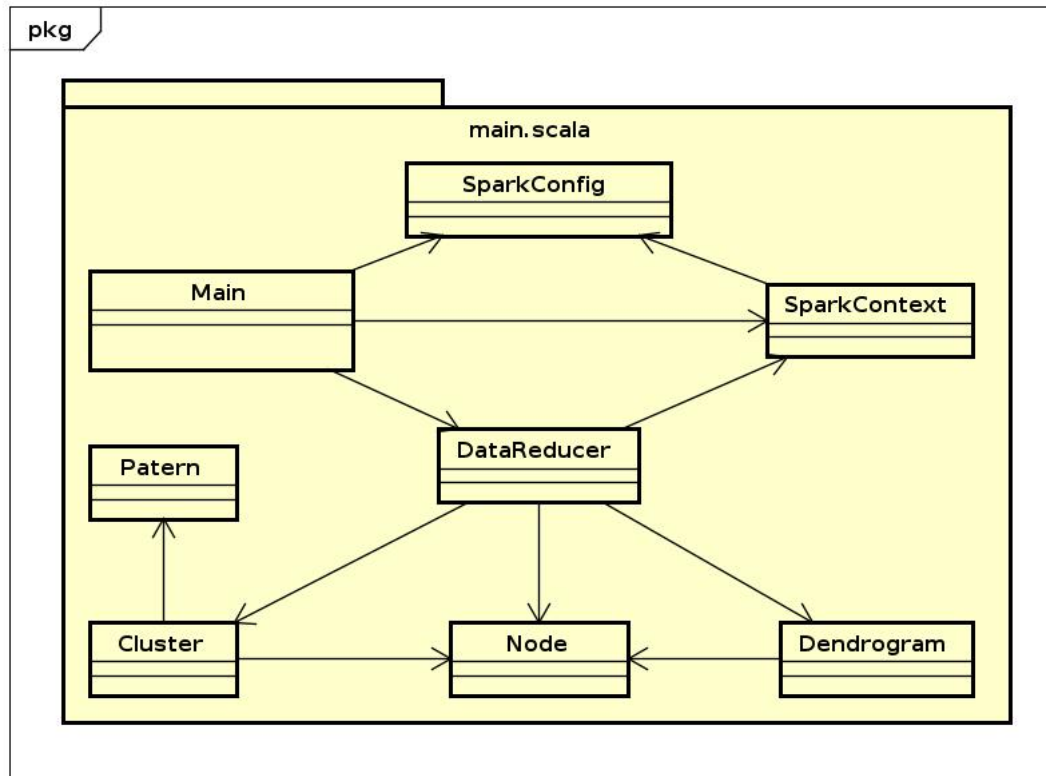
- Aktor: Pengguna
- Pre-kondisi: data yang akan diolah dimasukkan kepada HDFS.
- Pra-kondisi: hasil reduksi disimpan pada HDFS.
- Deskripsi: Fitur untuk menjalankan program untuk mereduksi data.
- Langkah-langkah:
  - (a) Pengguna mengisi JAR *path*, *input path*, dan *output path*.
  - (b) Pengguna mengisi jumlah *executor* dan besar *executor memory*.
  - (c) Pengguna mengisi jumlah partisi, batas maksimum objek, tipe metode, dan *cut-off distance*.
  - (d) Pengguna menekan tombol *submit*.
  - (e) Sistem melakukan pengolahan data dengan algoritma *Hierarchical Agglomerative Clustering* pada *cluster Hadoop*.
  - (f) Sistem membuka halaman baru untuk melihat tahap dan progres program.
  - (g) Sistem menyimpan hasil reduksi pada HDFS.

2. Nama *use case*: Mengunduh data

- Aktor: Pengguna
- Pre-kondisi: data yang akan diunduh sudah disimpan pada HDFS.
- Pra-kondisi: data dapat diunduh dari HDFS.
- Deskripsi: fitur untuk mengunduh data hasil reduksi.
- Langkah-langkah:
  - (a) Pengguna mengisi *path* dimana data disimpan pada HDFS.
  - (b) Sistem membuka halaman baru dimana pengguna dapat mengunduh data dari HDFS .

### 4.2.2 Diagram Kelas

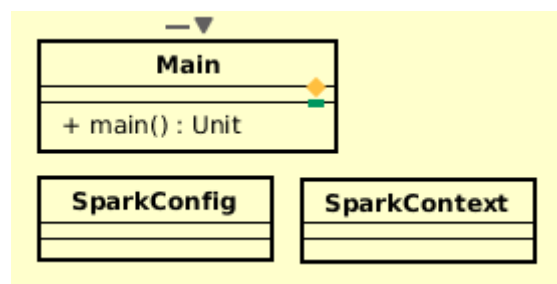
Berdasarkan hasil analisis perangkat lunak, dapat dirancang sebuah diagram kelas yang dapat dilihat pada Gambar 4.13.



Gambar 4.13: Diagram kelas

Berikut merupakan penjelasan detail mengenai kelas-kelas yang terdapat pada diagram kelas di atas:

- **Main, Spark Config, dan Spark Context**



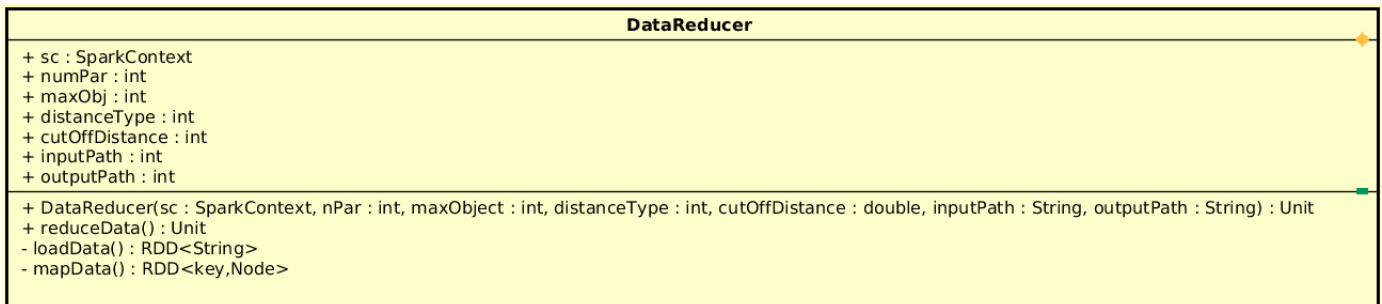
Gambar 4.14: Kelas Main, SparkConfig, SparkContext

Berikut adalah penjelasan dari ketiga kelas pada Gambar 4.14:

- *Main*: kelas *Main* memiliki *method main* yang merupakan titik masuk dari program. *Method* ini merupakan *method* pertama yang akan dieksekusi ketika program dijalankan.

- *SparkConfig*: kelas *SparkConfig* digunakan untuk mengatur konfigurasi untuk Spark. Pengaturan nama aplikasi, jumlah *core*, besar *memory*, dan lainnya dapat diatur pada kelas ini.
- *SparkContext*: kelas ini merupakan titik masuk untuk layanan-layanan dari Apache Spark.

#### • DataReducer

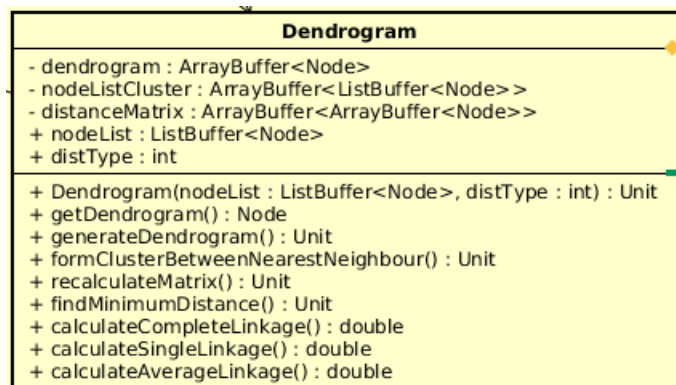


Gambar 4.15: Kelas DataReducer

Kelas *DataReducer* dirancang untuk memproses data. Proses reduksi secara paralel dilakukan pada kelas ini. Proses pemuatan dan penyimpanan data dilakukan pada kelas ini. Berdasarkan Gambar 4.15, berikut adalah penjelasan dari *methods* pada kelas *DataReducer*:

- *loadData*: *method* untuk memuat data berdasarkan *input path* yang diberikan.
- *mapData*: *method* untuk mengubah baris-baris atribut bertipe *String* menjadi objek *Node*. Setiap *Node* akan dipasangkan dengan *key* yaitu sebuah bilangan acak. *Method* ini akan mengembalikan RDD bertipe *<key,Node>*.
- *reduceData*: *method* untuk mereduksi data menggunakan *agglomerative clustering*. *Method* ini akan mengembalikan pola-pola dari setiap *clusters*.

#### • Dendrogram

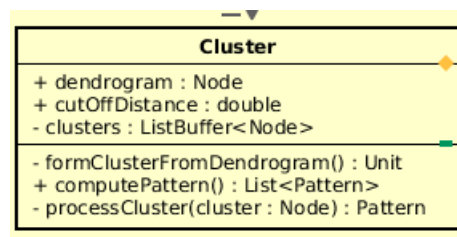


Gambar 4.16: Kelas Dendrogram

Kelas *Dendrogram* dirancang untuk memproses data dan membangun *dendrogram* sesuai algoritma *Hierarchical Agglomerative Clustering*. Berdasarkan Gambar 4.16, berikut adalah penjelasan *methods* pada kelas *Dendrogram*:

- *getDendrogram*: *method* ini mengembalikan *dendrogram*.
- *generateDendrogram*: *Method* untuk membangun *dendrogram* berdasarkan algoritma *Hierarchical Agglomerative Clustering*.
- *formClusterBetweenNearestNeighbour*: *method* untuk menggabungkan *cluster* terdekat.
- *recalculateMatrix*: *method* untuk menghitung ulang matriks jarak.
- *findMinimumDistance*: *method* untuk mencari jarak minimum antara dua *cluster*.
- *calculateCentroidLinkage*: *method* untuk mencari jarak antara *centorid* dua buah *cluster*.
- *calculateSingleLinkage*: *method* untuk mencari jarak minimum antara dua buah *cluster*.
- *calculateCompleteLinkage*: *method* untuk mencari jarak maksimum antara dua *cluster*.
- *calculateDistance*: *method* untuk mencari jarak antara dua buah *Node* berdasarkan atributnya.

#### • Cluster



Gambar 4.17: Kelas Cluster

Kelas *Cluster* dirancang untuk mengolah *cluster* dan menghasilkan pola dengan memotong *cluster*. Berdasarkan Gambar 4.17, berikut adalah penjelasan *methods* pada kelas *Cluster*:

- *formClusterFromDendrogram*: *method* ini bertugas untuk memotong *dendrogram* menjadi beberapa *cluster*.
- *computePattern*: *method* untuk mengolah potongan-potongan *cluster* menjadi pola dengan memanggil *method* *processCluster*.
- *processCluster*: *method* untuk memproses *cluster* dan membuat pola berdasarkan anggota-anggota pada *cluster*.

#### • Pattern

Pattern
+ max : double[] + min : double[] + average : double[] + SD : double[] + objCount : int
+ Pattern(max : double[], min : double[], average : double[], SD : double[], objCount : int) : Unit + getMaxArr() : double[] + getMinArr() : double[] + getAvgArr() : double[] + getObjCount() : int + getSD() : double[]

Gambar 4.18: Kelas Pattern

Kelas *Pattern* dirancang untuk merepresentasikan pola pada *cluster*. Berdasarkan Gambar 4.17, berikut adalah penjelasan *methods* pada kelas *Pattern*:

- *getMaxArr*: *method* ini mengembalikan *array* berisi nilai maksimum dari setiap atribut.
- *getMinArr*: *method* ini mengembalikan *array* berisi nilai minimum dari setiap atribut.
- *getAvgArr*: *method* ini mengembalikan *array* berisi nilai rata-rata dari setiap atribut.
- *getSDArr*: *method* ini mengembalikan *array* berisi nilai standar deviasi dari setiap atribut.
- *getObjCount*: *method* ini mengembalikan jumlah objek.

#### • Node

Node
- data : double[] - distance : double - rightNode : Node - leftNode : Node
+ setData(data : double[]) : Unit + setDistance(distance : double) + setRightNode(node : Node) : Unit + setLeftNode(node : Node) : Unit + getData() : double[] + getDistance() : double + getRightNode() : Node + getLeftNode() : Node

Gambar 4.19: Kelas Node

Kelas *Node* digunakan untuk membentuk pohon yang merepresentasikan *dendrogram*. Selain itu, kelas ini digunakan untuk merepresentasikan anggota pada *cluster*. Berdasarkan Gambar 4.19, berikut adalah penjelasan *methods* pada kelas *Node*:

- *setData*: *method* untuk memasukan nilai-nilai atribut.
- *setDistance*: *method* untuk megubah nilai jarak.
- *setRightNode*: *method* untuk menambahkan anak kanan *Node*.
- *setLeftNode*: *method* untuk menambahkan anak kiri *Node*.



- *getData: method* ini mengembalikan nilai-nilai atribut.
- *getDistance: method* ini mengembalikan jarak.
- *getRightNode: method* ini mengembalikan anak belah kanan dari *Node*.
- *getLeftNode: method* ini mengembalikan anak belah kiri dari *Node*.

### 4.2.3 Rancangan Antarmuka

Antarmuka dirancang untuk mempermudah pengguna dalam menjalankan program dan mengambil hasil data yang telah direduksi. Terdapat dua buah menu utama yang dapat dipilih oleh pengguna, menu Jalankan Program dan Lihat Pola. Menu Jalankan Program digunakan untuk menjalankan aplikasi dan menu Lihat Pola digunakan untuk mengunduh dan melihat hasil reduksi. Berikut adalah penjelasan rancangan antarmuka:

1. Perancangan halaman Jalankan Program untuk mempermudah pengguna menjalankan aplikasi. Pada halaman ini, disediakan *form* beserta *input* yang dibutuhkan untuk menjalankan aplikasi. Gambar rancangan antarmuka dapat dilihat pada Gambar 4.20

Gambar 4.20: Rancangan antarmuka menu Jalankan Program

Berdasarkan Gambar 4.20, berikut adalah penjelasan *input field* yang ada:

- *Spark JAR Path: field* untuk direktori JAR.
- *input path: field* untuk direktori file *input* pada HDFS.
- *output path: field* untuk direktori tempat penyimpanan hasil pada HDFS.
- *number of executor: field* untuk menentukan jumlah *executor* yang akan dipakai.
- *executor memory: field* untuk menentukan jumlah memori yang akan dipakai.
- *number of partition: field* untuk menentukan jumlah partisi untuk data.

- *max object: field* untuk membatasi jumlah objek pada yang akan diolah.
- *drop down (single linkage, complete linkage, centroid linkage):* kotak pilihan untuk memilih metode *single linkage, complete linkage* atau *centroid linkage* yang digunakan untuk memproses data.
- *cut off distance: field* untuk menentukan jarak untuk memotong *dendrogram* menjadi *clusters*.



## All Applications

Cluster

[About](#)  
[Nodes](#)  
[Node Labels](#)  
[Applications](#)  
[NEW](#)  
[NEW SAVING](#)  
[SUBMITTED](#)  
[ACCEPTED](#)  
[RUNNING](#)  
[FINISHED](#)  
[FAILED](#)  
[KILLED](#)  
[Scheduler](#)

Tools

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory
1	0	0	1	0	0 B	3 GB

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes
2	0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation
Capacity Scheduler	[MEMORY]	<memory:128, vCores:1>

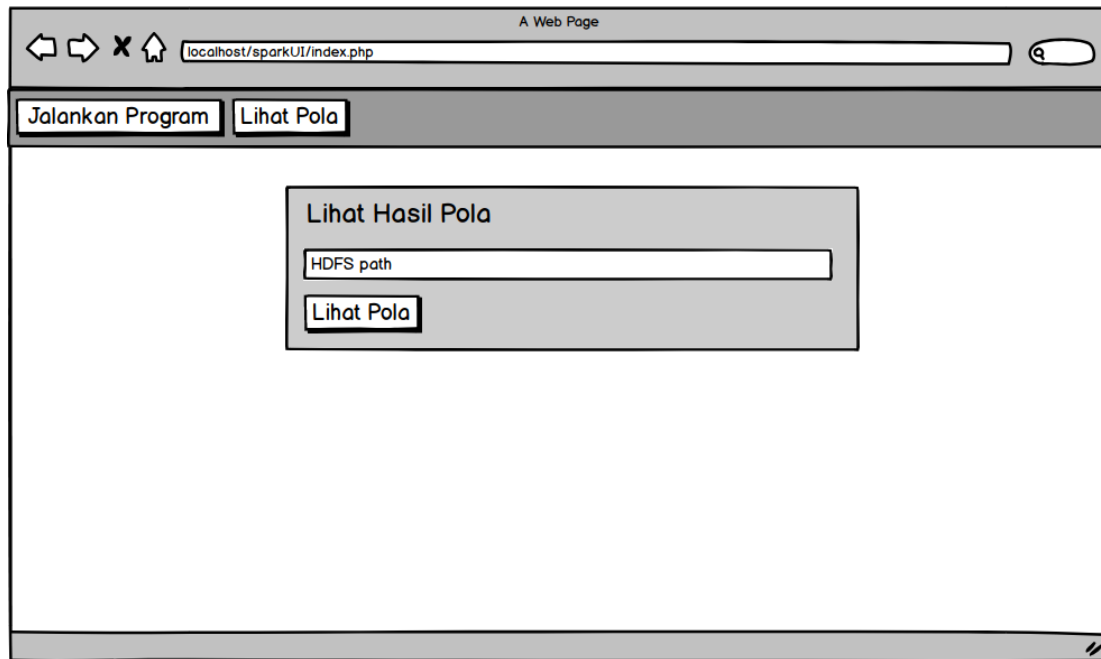
Show 20 entries

ID	User	Name	Application Type	Queue	Application Priority	StartTime	FinishTime	State	FinalStatus
<a href="#">application_1507721641254_0001</a>	hadoop	word count	MAPREDUCE	default	0	Wed Oct 11 13:39:33 +0200 2017	Wed Oct 11 13:40:13 +0200 2017	FINISHED	SUCCEEDED

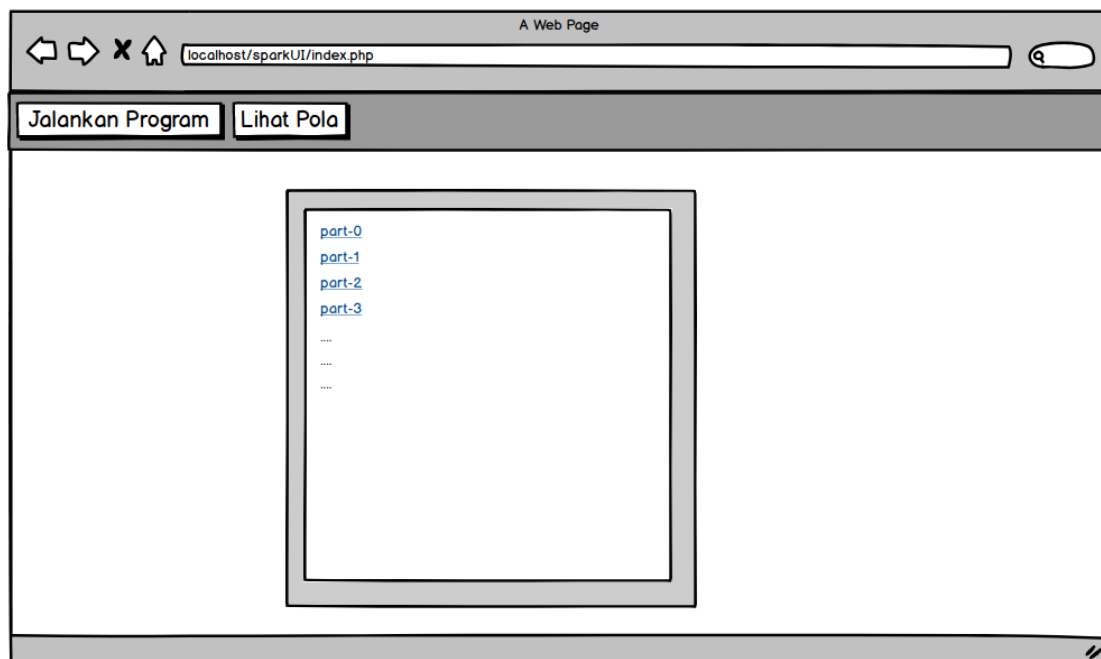
Showing 1 to 1 of 1 entries

Gambar 4.21: Halaman web Hadoop

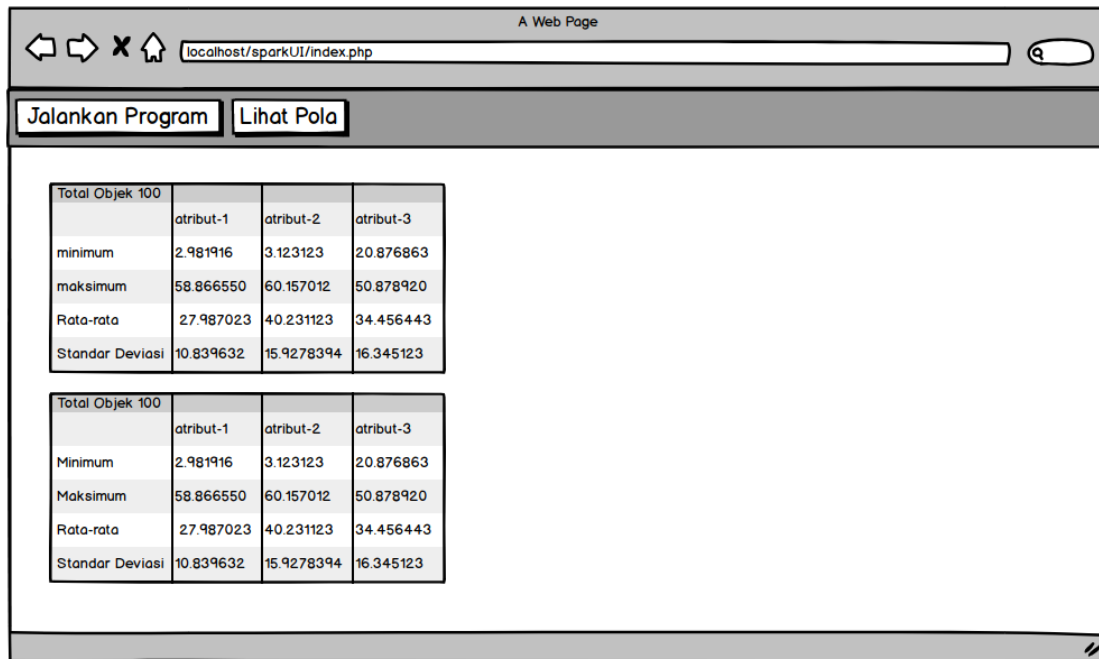
- Perancangan halaman antarmuka menu Lihat Pola (Gambar 4.22) digunakan untuk membuka direktori di mana data disimpan pada HDFS. Ketika pengguna memasukkan direktori, pengguna akan dipindahkan ke halaman baru (Gambar 4.23) dan sebuah halaman (Gambar 4.25) akan dibuka untuk menampilkan data yang dapat diunduh. Ketika pengguna menekan salah satu nama partisi pada halaman berikutnya (Gambar 4.23), pengguna dapat melihat pola-pola dari partisi tersebut di halaman (Gambar 4.24).



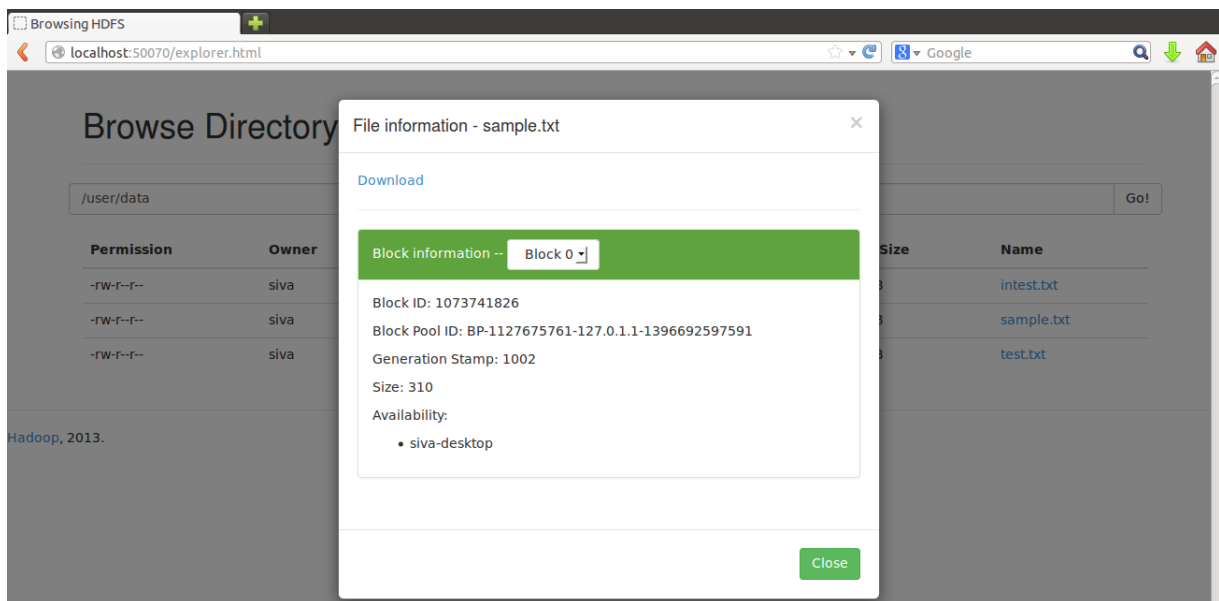
Gambar 4.22: Rancangan antarmuka menu Lihat Pola



Gambar 4.23: Rancangan antarmuka halaman partisi



Gambar 4.24: Rancangan antarmuka halaman pola



Gambar 4.25: Halaman web HDFS

## BAB 5

# IMPLEMENTASI DAN PENGUJIAN PERANGKAT LUNAK

## 5.1 Implementasi Perangkat Lunak

### 5.1.1 Lingkungan Perangkat Keras

Perangkat keras yang digunakan dalam membangun perangkat lunak adalah sebuah PC dengan spesifikasi berikut:

- *Processor*: Intel i7 4790K @4.00 GHz
- RAM: 16 GB DDR3
- VGA: NVIDIA GeForce GTX 750TI 2GB
- *Harddisk*: 1TB + 256GB SSD

### 5.1.2 Lingkungan Perangkat Lunak

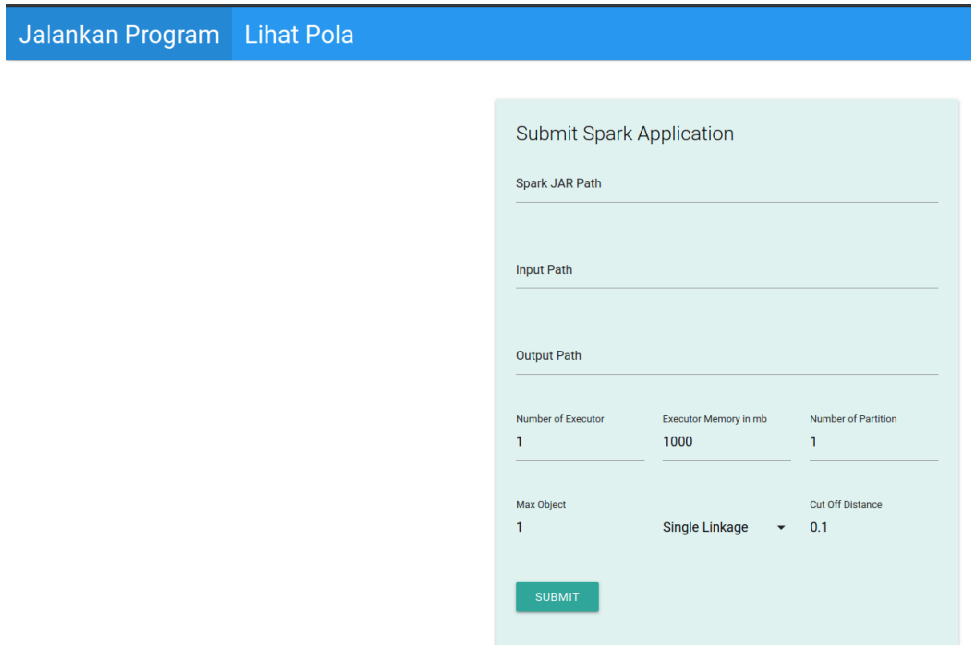
Perangkat lunak yang digunakan untuk membangun perangkat lunak adalah sebagai berikut:

- Sistem Operasi: Ubuntu 18.04.2 LTS
- Bahasa Pemrograman: Scala
- IDE: IntelliJ IDE 2018
- Versi Java: JDK 1.8.0\_181
- Versi Scala: Scala 2.11.12
- Versi SBT: SBT 1.2.8
- Library Dependency:
  - org.apache.spark:spark-core 2.1.0
  - org.scala-lang:scala-library 2.11.12

### 5.1.3 User Interface

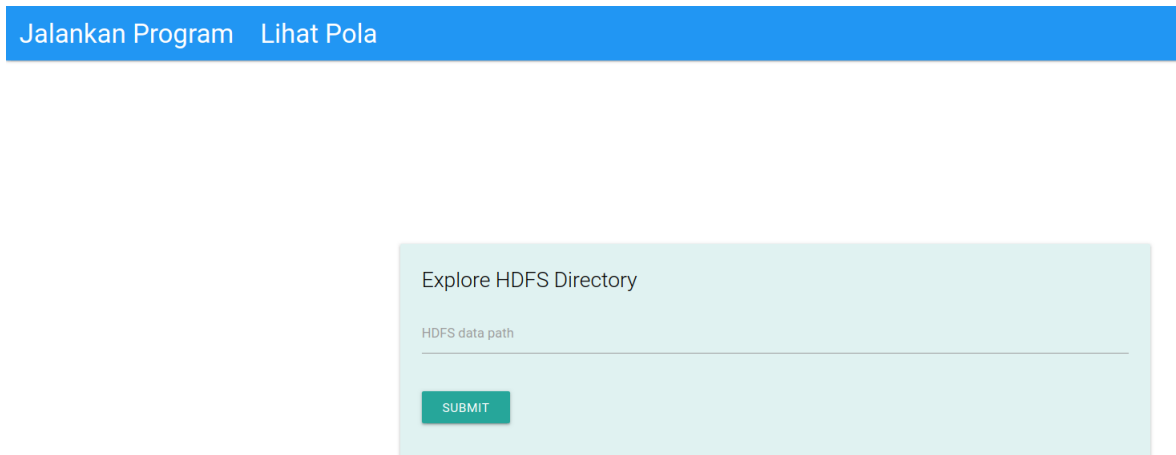
Implementasi rancangan tampilan antarmuka pada perangkat lunak ini menggunakan html,css, dan php. Berikut adalah tampilan setiap halaman:

1. Implementasi antarmuka untuk menu *Submit* dapat dilihat pada Gambar 5.1.



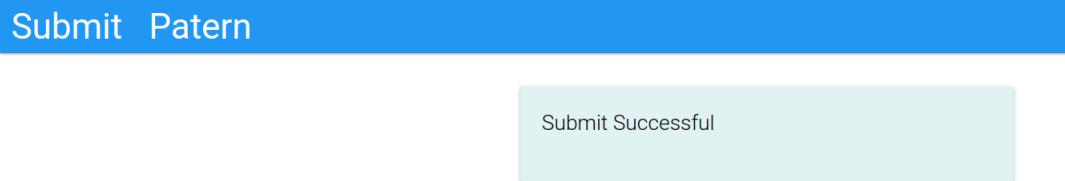
Gambar 5.1: Tampilan menu *submit*

2. Implementasi antarmuka untuk menu *Data* dapat dilihat pada Gambar 5.2.



Gambar 5.2: Tampilan menu *Data*

3. Implementasi antarmuka sesudah melakukan *submit* dapat dilihat pada Gambar 5.3.



Gambar 5.3: Tampilan halaman sesudah *submit*

- 1      4. Implementasi antarmuka halaman *list* dapat dilihat pada Gambar 5.4.



Gambar 5.4: Tampilan halaman *list*

- 2      5. Implementasi antarmuka halaman *data* dapat dilihat pada Gambar 5.5.

## Submit Patern

Total Obj = 13	attribute-1	attribute-2
Minimum	2.989196145255757	32.668909774235246
Maximum	58.86655003746121	90.87631433387419
Average	32.87713041936097	65.42538643824919
Stardard Deviation	14.941607927042684	17.27969410300335

Total Obj = 5	attribute-1	attribute-2
Minimum	1.715184579886253	6.548756714677017
Maximum	25.99400990511951	26.14315566259172
Average	16.39405819212848	17.238589783987994
Stardard Deviation	11.24161321121706	7.833594989767486

Total Obj = 5	attribute-1	attribute-2
Minimum	48.66269805215405	5.134186149105857

Gambar 5.5: Tampilan halaman data

## 5.2 Pengujian Fungsional Perangkat Lunak

Perangkat lunak yang disusun oleh penulis telah diuji untuk membuktikan kebenaran dari perangkat lunak. Program akan dieksekusi dan kemudian diamati apakah hasil sesuai dengan yang diinginkan. Perangkat lunak akan diberikan data dengan ukuran yang kecil berserta *parameter* yang sudah ditentukan.

- Pada percobaan pertama, akan digunakan metode *single linkage*, dengan jumlah partisi = 1, jumlah objek maksimum = 4, dan nilai *cut-off distance* = 0.8. Berikut adalah data yang digunakan untuk pengujian:

4.0, 5.0



1 3.0,7.0  
 2 4.0,3.0  
 3 10.0,7.0  
 4 10.0,10.0

5 Hasil dari percobaan pertama adalah sebagai berikut:

6 3  
 7 3.0,3.0  
 8 4.0,7.0  
 9 3.666666666666665,5.0  
 10 0.5773502691896258,2.0  
 11 1  
 12 10.0,7.0  
 13 10.0,7.0  
 14 10.0,7.0  
 15 0.0,0.0  
 16 1  
 17 10.0,10.0  
 18 10.0,10.0  
 19 10.0,10.0  
 20 0.0,0.0

21 • Pada percobaan kedua, akan digunakan metode *complete linkage*, dengan jumlah partisi = 1, jumlah  
 22 objek maksimum = 4, dan nilai *cut-off distance* = 0.8. Berikut adalah data yang digunakan untuk  
 23 pengujian:

24 4.0,5.0  
 25 3.0,7.0  
 26 4.0,3.0  
 27 10.0,7.0  
 28 10.0,10.0

29 Hasil dari percobaan kedua adalah sebagai berikut:

30 3  
 31 3.0,3.0  
 32 4.0,7.0  
 33 3.666666666666665,5.0  
 34 0.5773502691896258,2.0  
 35 1  
 36 10.0,7.0  
 37 10.0,7.0

10.0,7.0  
 0.0,0.0  
 1  
 10.0,10.0  
 10.0,10.0  
 10.0,10.0  
 0.0,0.0

- Pada percobaan ketiga, akan digunakan metode *centroid linkage*, dengan jumlah partisi = 1, jumlah objek maksimum = 4, dan nilai *cut-off distance* = 0.8. Berikut adalah data yang digunakan untuk pengujian:

4.0,5.0  
 3.0,7.0  
 4.0,3.0  
 10.0,7.0  
 10.0,10.0

Hasil dari percobaan ketiga adalah sebagai berikut:

3  
 3.0,3.0  
 4.0,7.0  
 3.6666666666666665,5.0  
 0.5773502691896258,2.0  
 1  
 10.0,7.0  
 10.0,7.0  
 10.0,7.0  
 0.0,0.0  
 1  
 10.0,10.0  
 10.0,10.0  
 10.0,10.0  
 0.0,0.0

Berdasarkan hasil ketiga percobaan yang didapat, maka dapat disimpulkan bahwa perangkat lunak sudah dapat melakukan proses reduksi data menggunakan algoritma *Agglomerative Clustering* berdasarkan metode yang dipilih dengan benar. Pola yang dihasilkan oleh perangkat lunak sudah sesuai dengan apa yang diharapkan.

### 5.3 Hasil Eksperimen Perangkat Lunak

Pada bagian ini akan diuji performa perangkat lunak Spark dan Hadoop. Kedua perangkat lunak akan dibandingkan hasil eksekusi waktunya. Karena perangkat lunak hadoop tidak dapat menghitung standar deviasi,

maka perangkat lunak Hadoop akan dibandingkan dengan perangkat lunak Spark yang tidak menghitung standar deviasi dan yang menghitung standar deviasi. Data yang digunakan pada percobaan merupakan data yang dihasilkan secara acak dengan ukuran yang berbeda-beda. Data-data tersebut memiliki dua atribut bilangan pecahan yang dipisahkan dengan tanda koma. Jumlah objek pada setiap ukuran data dapat dilihat pada Tabel 5.1.

Tabel 5.1: Tabel data yang digunakan pada eksperimen

Ukuran Data	Jumlah Objek	Jumlah Block
1 GB	36000000	40
2 GB	64000000	70
3 GB	81000000	89
5 GB	144000000	157
10 GB	256000000	279
15 GB	400000000	435
20 GB	529000000	576

Berikut adalah spesifikasi perangkat keras yang digunakan:

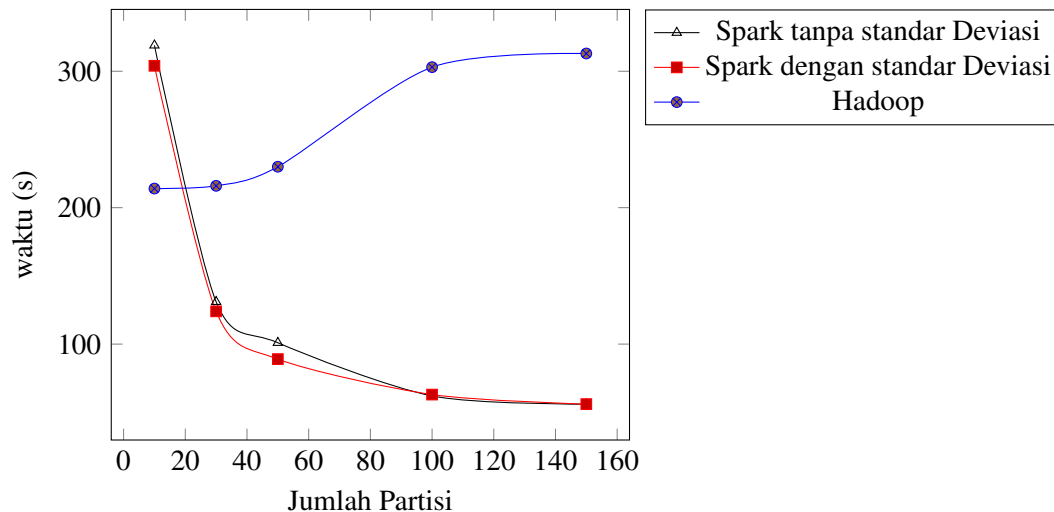
- *Processor*: Intel core i5 8500 @3.00 GHz, 6 core
- RAM: 8GB
- *Harddisk*: 500GB
- Sistem Operasi: Ubuntu 18.0.4

## 5.4 Percobaan Dampak Partisi pada Performa Perangkat Lunak Spark dan Hadoop

Pada percobaan ini akan dilihat waktu eksekusi Spark dan Hadoop berdasarkan jumlah partisi yang berbeda. Percobaan ini akan menggunakan 1 komputer sebagai komputer *master* dan 10 komputer lainnya sebagai *worker* dengan setiap *worker* menggunakan 1 core. Ukuran data yang digunakan adalah 1 GB. Metode yang digunakan adalah metode *single linkage*, dengan nilai *cut-off distance* adalah 0,8 dan jumlah objek maksimum untuk setiap *dendrogram* adalah 30. Tabel (5.2) berikut adalah hasil dari eksperimen:

Tabel 5.2: Percobaan Jumlah Partisi Hadoop dan Spark dengan Ukuran Data 1 GB

Ukuran Data (GB)	Jumlah Partisi	Waktu Eksekusi Spark Tanpa standar Deviasi (Detik)	Waktu Eksekusi Spark (Detik)	Waktu Eksekusi Hadoop (Detik)	Hasil Reduksi Spark Tanpa standar Deviasi (GB)	Hasil Reduksi Spark (GB)	Hasil Reduksi Hadoop (GB)
1	10	319	304	214	0.54	0.67	0.57
1	30	131	124	216	0.54	0.67	0.57
1	50	101	89	230	0.54	0.67	0.57
1	100	62	63	303	0.54	0.67	0.57
1	150	56	56	313	0.54	0.67	0.57



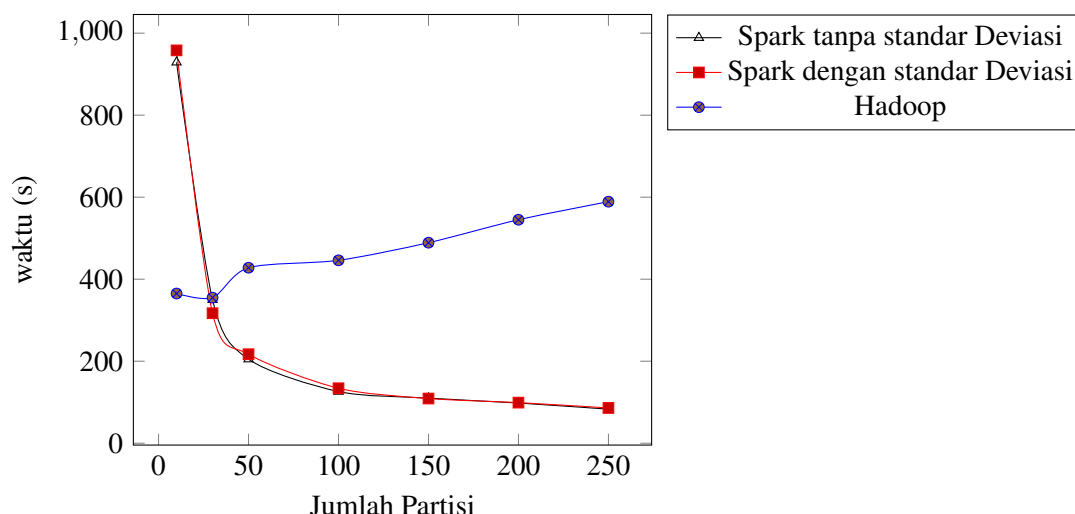
Gambar 5.6: dengan ukuran data 1GB, jumlah objek maksimum 30, dan total 10 core

Berdasarkan hasil grafik ( 5.6), dapat dilihat bahwa waktu eksekusi Spark menurun dan waktu eksekusi Hadoop meningkat ketika jumlah partisi diperbesar. Waktu eksekusi Hadoop menaik secara konsisten ketika jumlah partisi diperbesarkan. Waktu eksekusi Spark menurun drastis pada awalnya ketika jumlah partisi ditingkatkan sampai titik tertentu dimana peningkatan jumlah partisi tidak memiliki dampak yang sangat drastis pada waktu eksekusi Spark. Tidak ada perbedaan yang jauh antara waktu eksekusi aplikasi Spark dengan standar deviasi maupun yang tidak. Aplikasi Spark memiliki waktu eksekusi yang lebih baik dibanding Hadoop pada jumlah partisi yang besar dan waktu eksekusi yang lebih buruk pada jumlah partisi yang kecil. Waktu eksekusi Spark pada partisi yang besar lebih cepat dibanding waktu eksekusi Hadoop terkecil. Aplikasi Spark lebih cepat dibanding Hadoop asalakan jumlah partisi diatur dengan benar.

Pada percobaan ini akan dilihat waktu eksekusi Spark dan Hadoop berdasarkan jumlah partisi yang berbeda. Percobaan ini akan menggunakan 1 komputer sebagai komputer master dan 10 komputer lainnya sebagai worker dengan setiap worker menggunakan 1 core. Ukuran data yang digunakan adalah 2 GB. Metode yang digunakan adalah metode *single linkage*, dengan nilai *cut-off distance* adalah 0,8 dan jumlah objek maksimum untuk setiap *dendrogram* adalah 30. Tabel ( 5.3) berikut adalah hasil dari eksperimen:

Tabel 5.3: Percobaan Jumlah Partisi Hadoop dan Spark dengan Ukuran Data 2 GB

Ukuran Data (GB)	Jumlah Partisi	Waktu Eksekusi Spark Tanpa standar Deviasi (Detik)	Waktu Eksekusi Spark (Detik)	Waktu Eksekusi Hadoop (Detik)	Hasil Reduksi Spark Tanpa standar Deviasi (GB)	Hasil Reduksi Spark (GB)	Hasil Reduksi Hadoop (GB)
2	10	929	958	365	0.96	1.2	1
2	30	350	317	355	0.96	1.2	1
2	50	205	217	428	0.96	1.2	1
2	100	126	134	446	0.96	1.2	1
2	150	110	109	489	0.96	1.2	1
2	200	98	99	545	0.96	1.2	1
2	250	83	86	589	0.96	1.2	1



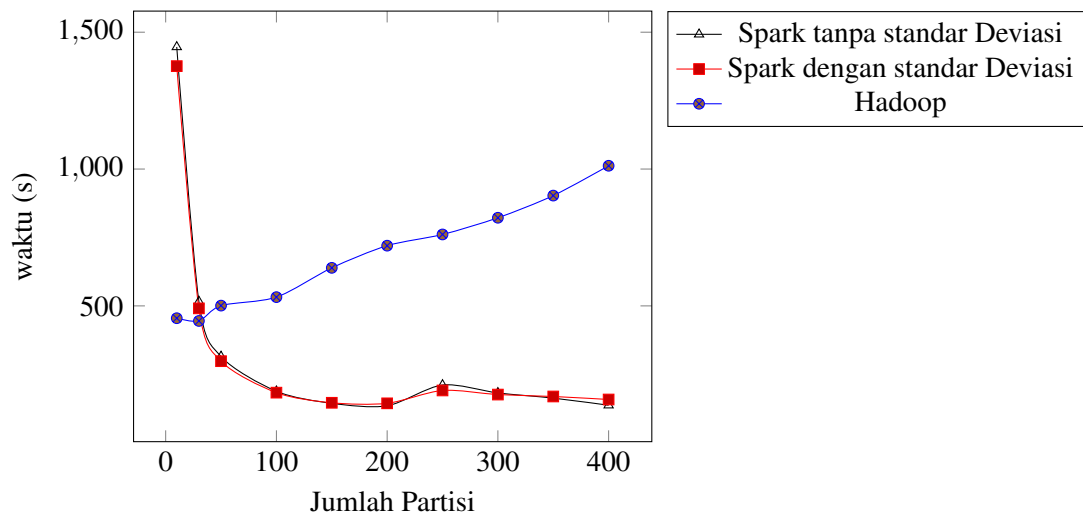
Gambar 5.7: Hasil Percobaan Partisi Spark dan Hadoop dengan ukuran data 2GB, jumlah objek maksimum 30, dan total 10 core

Berdasarkan hasil grafik ( 5.7), dapat dilihat bahwa waktu eksekusi Spark menurun dan waktu eksekusi Hadoop meningkat ketika jumlah partisi diperbesar. Waktu eksekusi Hadoop menaik secara konsisten ketika jumlah partisi diperbesarkan. Waktu eksekusi Spark menurun drastis pada awalnya ketika jumlah partisi ditingkatkan sampai titik tertentu dimana peningkatan jumlah partisi tidak memiliki dampak yang sangat drastis pada waktu eksekusi Spark. Tidak ada perbedaan yang jauh antara waktu eksekusi aplikasi Spark dengan standar deviasi maupun yang tidak. Aplikasi Spark memiliki waktu eksekusi yang lebih baik dibanding Hadoop pada jumlah partisi yang besar dan waktu eksekusi yang lebih buruk pada jumlah partisi yang kecil. Waktu eksekusi Spark pada partisi yang besar lebih cepat dibanding waktu eksekusi Hadoop terkecil. Aplikasi Spark lebih cepat dibanding Hadoop asalakan jumlah partisi diatur dengan benar.

Pada percobaan ini akan dilihat waktu eksekusi Spark dan Hadoop berdasarkan jumlah partisi yang berbeda. Percobaan ini akan menggunakan 1 komputer sebagai komputer master dan 10 komputer lainnya sebagai worker dengan setiap worker menggunakan 1 core. Ukuran data yang digunakan adalah 3 GB. Metode yang digunakan adalah metode *single linkage*, dengan nilai *cut-off distance* adalah 0,8 dan jumlah objek maksimum untuk setiap *dendrogram* adalah 30. Tabel ( 5.4) berikut adalah hasil dari eksperimen:

Tabel 5.4: Percobaan Jumlah Partisi Hadoop dan Spark dengan Ukuran Data 3 GB

Ukuran Data (GB)	Jumlah Partisi	Waktu Eksekusi Spark Tanpa standar Deviasi (Detik)	Waktu Eksekusi Spark (Detik)	Waktu Eksekusi Hadoop (Detik)	Hasil Reduksi Spark Tanpa standar Deviasi (GB)	Hasil Reduksi Spark (GB)	Hasil Reduksi Hadoop (GB)
3	10	1446	1376	455	1.2	1.5	1.2
3	30	516	491	445	1.2	1.5	1.2
3	50	315	298	501	1.2	1.5	1.2
3	100	188	183	532	1.2	1.5	1.2
3	150	144	146	639	1.2	1.5	1.2
3	200	135	144	720	1.2	1.5	1.2
3	250	211	191	761	1.2	1.5	1.2
3	300	182	176	822	1.2	1.5	1.2
3	350	163	169	903	1.2	1.5	1.2
3	400	137	158	1012	1.2	1.5	1.2



Gambar 5.8: Hasil Percobaan Partisi Spark dan Hadoop dengan ukuran data 3GB, Objek Maksimum 30, dan Total 10 Core

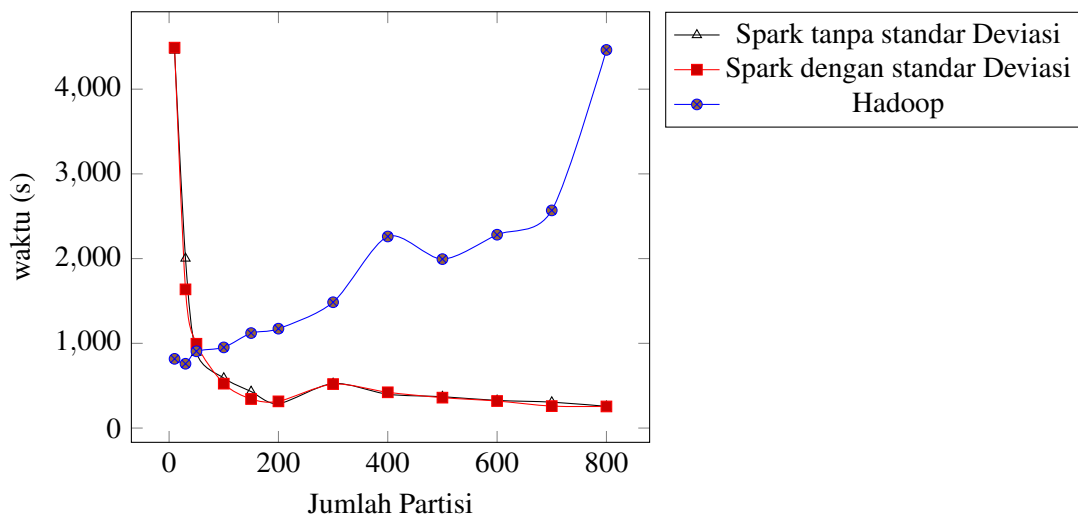
Berdasarkan hasil grafik ( 5.8), dapat dilihat bahwa waktu eksekusi Spark menurun dan waktu eksekusi Hadoop meningkat ketika jumlah partisi diperbesar. Waktu eksekusi Hadoop menaik secara konsisten ketika jumlah partisi diperbesarkan. Waktu eksekusi Spark menurun drastis pada awalnya ketika jumlah partisi ditingkatkan sampai titik tertentu dimana peningkatan jumlah partisi tidak memiliki dampak yang sangat drastis pada waktu eksekusi Spark. Tidak ada perbedaan yang jauh antara waktu eksekusi aplikasi Spark dengan standar deviasi maupun yang tidak. Aplikasi Spark memiliki waktu eksekusi yang lebih baik dibanding Hadoop pada jumlah partisi yang besar dan waktu eksekusi yang lebih buruk pada jumlah partisi yang kecil. Waktu eksekusi Spark pada partisi yang besar lebih cepat dibanding waktu eksekusi Hadoop terkecil. Aplikasi Spark lebih cepat dibanding Hadoop asalakan jumlah partisi diatur dengan benar.

Pada percobaan ini akan dilihat waktu eksekusi Spark dan Hadoop berdasarkan jumlah partisi yang berbeda. Percobaan ini akan menggunakan 1 komputer sebagai komputer master dan 10 komputer lainnya sebagai worker dengan setiap worker menggunakan 1 core. Ukuran data yang digunakan adalah 5 GB.

- 1 Metode yang digunakan adalah metode *single linkage*, dengan nilai *cut-off distance* adalah 0,8 dan jumlah
- 2 objek maksimum untuk setiap *dendrogram* adalah 30. Tabel ( 5.5) berikut adalah hasil dari eksperimen:

Tabel 5.5: Percobaan Jumlah Partisi Hadoop dan Spark dengan Ukuran Data 5 GB

Ukuran Data (GB)	Jumlah Partisi	Waktu Eksekusi Spark Tanpa standar Deviasi (Detik)	Waktu Eksekusi Spark (Detik)	Waktu Eksekusi Hadoop (Detik)	Hasil Reduksi Spark Tanpa standar Deviasi (GB)	Hasil Reduksi Spark (GB)	Hasil Reduksi Hadoop (GB)
5	10	4490	4457	817	2.1	2.6	2.2
5	30	1637	2002	759	2.1	2.6	2.2
5	50	995	891	906	2.1	2.6	2.2
5	100	524	590	952	2.1	2.6	2.2
5	150	343	431	1121	2.1	2.6	2.2
5	200	315	288	1173	2.1	2.6	2.2
5	300	519	526	1485	2.1	2.6	2.2
5	400	422	399	2261	2.1	2.6	2.2
5	500	359	370	1994	2.1	2.6	2.2
5	600	319	326	2282	2.1	2.6	2.2
5	700	259	306	2569	2.1	2.6	2.2
5	800	255	256	4463	2.1	2.6	2.2



Gambar 5.9: Hasil Percobaan Partisi Spark dan Hadoop dengan ukuran data 5GB, Objek Maksimum 30, dan Total 10 Core

Berdasarkan hasil grafik ( 5.9), dapat dilihat bahwa waktu eksekusi Spark menurun dan waktu eksekusi Hadoop meningkat ketika jumlah partisi diperbesar. Waktu eksekusi Hadoop menaik secara konsisten ketika jumlah partisi diperbesarkan. Waktu eksekusi Spark menurun drastis pada awalnya ketika jumlah partisi ditingkatkan sampai titik tertentu dimana peningkatan jumlah partisi tidak memiliki dampak yang sangat drastis pada waktu eksekusi Spark. Tidak ada perbedaan yang jauh antara waktu eksekusi aplikasi Spark dengan standar deviasi maupun yang tidak. Aplikasi Spark memiliki waktu eksekusi yang lebih baik dibanding Hadoop pada jumlah partisi yang besar dan waktu eksekusi yang lebih buruk pada jumlah partisi yang kecil. Waktu eksekusi Spark pada partisi yang besar lebih cepat dibanding waktu eksekusi Hadoop terkecil. Aplikasi Spark lebih cepat dibanding Hadoop asalakan jumlah partisi diatur dengan benar.

Pada percobaan ini akan dilihat waktu eksekusi Spark dan Hadoop berdasarkan jumlah partisi yang berbeda. Percobaan ini akan menggunakan 1 komputer sebagai komputer master dan 10 komputer lainnya sebagai worker dengan setiap worker menggunakan 1 core. Ukuran data yang digunakan adalah 10 GB. Metode yang digunakan adalah metode *single linkage*, dengan nilai *cut-off distance* adalah 0,8 dan jumlah objek maksimum untuk setiap *dendrogram* adalah 30. Tabel 5.6 dan Tabel 5.8 berikut adalah hasil dari eksperimen:

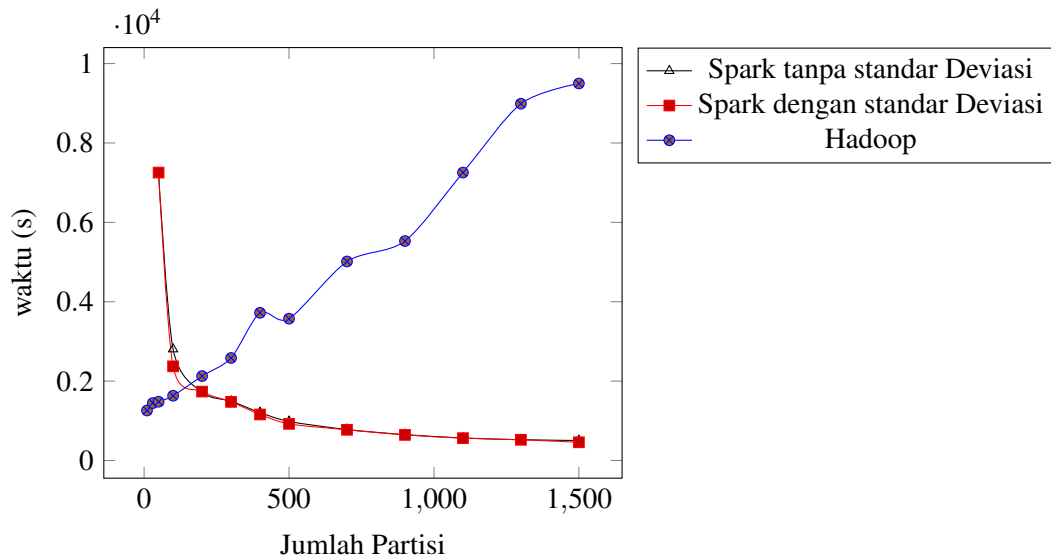
Tabel 5.6: Percobaan Jumlah Partisi Spark dengan Ukuran Data 10 GB

Ukuran Data (GB)	Jumlah Partisi	Waktu Eksekusi Spark Tanpa standar Deviasi (Detik)	Waktu Eksekusi Spark (Detik)	Hasil Reduksi Spark standar Deviasi (GB)	Hasil Reduksi Spark (GB)
10	50	7254	7236	3.7	4.6
10	100	237	2805	3.7	4.6
10	200	1736	1718	3.7	4.6
10	300	1477	1494	3.7	4.6
10	400	1160	1207	3.7	4.6
10	500	923	984	3.7	4.6
10	600	774	780	3.7	4.6
10	700	645	652	3.7	4.6
10	900	563	568	3.7	4.6
10	1100	522	524	3.7	4.6
10	1300	359	504	3.7	4.6
10	1500	255	256	3.7	4.6

Tabel 5.7: Percobaan Jumlah Partisi Hadoop dengan Ukuran Data 10 GB

Ukuran Data (GB)	Jumlah Partisi	Waktu Eksekusi Hadoop (Detik)	Hasil Reduksi Hadoop (GB)
10	10	1260	3.9
10	30	1446	3.9
10	50	1481	3.9
10	100	1631	3.9
10	200	2127	3.9
10	300	2583	3.9
10	400	3721	3.9
10	500	3573	3.9
10	700	5014	3.9
10	900	5529	3.9
10	1100	7254	3.9
10	1300	8989	3.9
10	1500	9499	3.9





Gambar 5.10: Hasil Percobaan Partisi Spark dan Hadoop dengan ukuran data 10GB, Objek Maksimum 30, dan Total 10 Core

Berdasarkan hasil grafik ( 5.10), dapat dilihat bahwa waktu eksekusi Spark menurun dan waktu eksekusi Hadoop meningkat ketika jumlah partisi diperbesar. Waktu eksekusi Hadoop menaik secara konsisten ketika jumlah partisi diperbesarkan. Waktu eksekusi Spark menurun drastis pada awalnya ketika jumlah partisi ditingkatkan sampai titik tertentu dimana peningkatan jumlah partisi tidak memiliki dampak yang sangat drastis pada waktu eksekusi Spark. Tidak ada perbedaan yang jauh antara waktu eksekusi aplikasi Spark dengan standar deviasi maupun yang tidak. Aplikasi Spark memiliki waktu eksekusi yang lebih baik dibanding Hadoop pada jumlah partisi yang besar dan waktu eksekusi yang lebih buruk pada jumlah partisi yang kecil. Waktu eksekusi Spark pada partisi yang besar lebih cepat dibanding waktu eksekusi Hadoop terkecil. Aplikasi Spark lebih cepat dibanding Hadoop asalakan jumlah partisi diatur dengan benar.

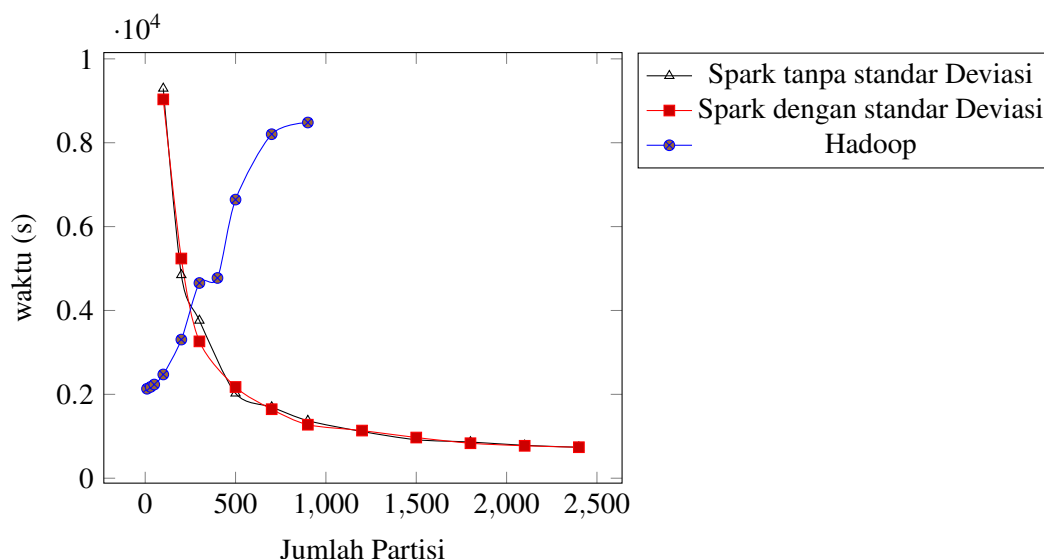
Pada percobaan ini akan dilihat waktu eksekusi Spark dan Hadoop berdasarkan jumlah partisi yang berbeda. Percobaan ini akan menggunakan 1 komputer sebagai komputer master dan 10 komputer lainnya sebagai worker dengan setiap worker menggunakan 1 core. Ukuran data yang digunakan adalah 15 GB. Metode yang digunakan adalah metode *single linkage*, dengan nilai *cut-off distance* adalah 0,8 dan jumlah objek maksimum untuk setiap *dendrogram* adalah 30. Tabel ( 5.8 dan Tabel ( 5.9) berikut adalah hasil dari eksperimen:

Tabel 5.8: Percobaan Jumlah Partisi Spark dengan Ukuran Data 15 GB

Ukuran Data )	Jumlah Partisi	Waktu Eksekusi Spark Tanpa standar Deviasi (GB)	Waktu Eksekusi Spark (detik)	Hasil Reduksi Spark Tanpa standar Deviasi )	Hasil Reduksi Spark )
15	100	9034	9294	5.8	7.3
15	200	5239	4847	5.8	7.3
15	300	3263	3761	5.8	7.3
15	500	2175	2024	5.8	7.3
15	700	1645	1696	5.8	7.3
15	900	1276	1372	5.8	7.3
15	1200	1136	1114	5.8	7.3
15	1500	970	918	5.8	7.3
15	1800	834	863	5.8	7.3
15	2100	773	783	5.8	7.3
15	2400	739	738	5.8	7.3

Tabel 5.9: Percobaan Jumlah Partisi Hadoop dengan Ukuran Data 15 GB

Ukuran Data (GB)	Jumlah Partisi	Waktu Eksekusi Hadoop (detik)	Hasil Reduksi Hadoop (GB)
15	10	2133	3.9
15	30	2177	3.9
15	50	2234	3.9
15	100	2474	3.9
15	200	3306	3.9
15	300	4655	3.9
15	400	4775	3.9
15	500	6644	3.9
15	700	8203	3.9
15	900	8482	3.9



Gambar 5.11: Hasil Percobaan Partisi Spark dan Hadoop dengan ukuran data 15GB, Objek Maksimum 30, dan Total 10 Core

Berdasarkan hasil grafik ( 5.11), dapat dilihat bahwa waktu eksekusi Spark menurun dan waktu eksekusi Hadoop meningkat ketika jumlah partisi diperbesar. Waktu eksekusi Hadoop menaik secara konsisten ketika jumlah partisi diperbesarkan. Waktu eksekusi Spark menurun drastis pada awalnya ketika jumlah partisi ditingkatkan sampai titik tertentu dimana peningkatan jumlah partisi tidak memiliki dampak yang sangat drastis pada waktu eksekusi Spark. Tidak ada perbedaan yang jauh antara waktu eksekusi aplikasi Spark dengan standar deviasi maupun yang tidak. Aplikasi Spark memiliki waktu eksekusi yang lebih baik dibanding Hadoop pada jumlah partisi yang besar dan waktu eksekusi yang lebih buruk pada jumlah partisi yang kecil. Waktu eksekusi Spark pada partisi yang besar lebih cepat dibanding waktu eksekusi Hadoop terkecil. Aplikasi Spark lebih cepat dibanding Hadoop asalakan jumlah partisi diatur dengan benar.

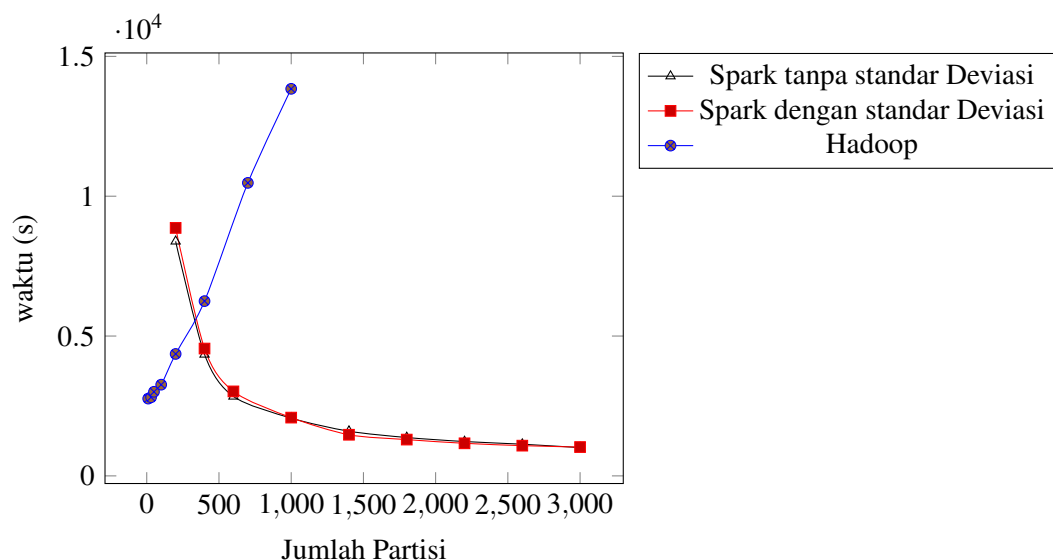
Pada percobaan ini akan dilihat waktu eksekusi Spark dan Hadoop berdasarkan jumlah partisi yang berbeda. Percobaan ini akan menggunakan 1 komputer sebagai komputer master dan 10 komputer lainnya sebagai worker dengan setiap worker menggunakan 1 core. Ukuran data yang digunakan adalah 20 GB. Metode yang digunakan adalah metode *single linkage*, dengan nilai *cut-off distance* adalah 0,8 dan jumlah objek maksimum untuk setiap *dendrogram* adalah 30. Tabel ( 5.10 dan Tabel ( 5.11) berikut adalah hasil dari eksperimen:

Tabel 5.10: Percobaan Jumlah Partisi Spark dengan Ukuran Data 20 GB

Ukuran Data (GB)	Jumlah Partisi	Waktu Eksekusi Spark standar (detik)	Eksekusi Tanpa Deviasi	Waktu Eksekusi Spark (detik)	Hasil Spark standar (GB)	Reduksi Tanpa Deviasi	Hasil Reduksi Spark (GB)
20	200	8866		8386	7.7		9.6
20	400	4553		4342	7.7		9.6
20	600	3021		2841	7.7		9.6
20	1000	2084		2065	7.7		9.6
20	1400	1471		1598	7.7		9.6
20	1800	1298		1372	7.7		9.6
20	2200	1165		1228	7.7		9.6
20	2600	1081		1133	7.7		9.6
20	3000	1031		1010	7.7		9.6

Tabel 5.11: Percobaan Jumlah Partisi Hadoop dengan Ukuran Data 20 GB

Ukuran Data (GB)	Jumlah Partisi	Waktu Eksekusi Hadoop (detik)	Hasil Reduksi Hadoop (GB)
20	10	2763	8.1
20	30	2811	8.1
20	50	3007	8.1
20	100	3261	8.1
20	200	4360	8.1
20	400	6249	8.1
20	700	10476	8.1
20	1000	13839	8.1



Gambar 5.12: Hasil Percobaan Partisi Spark dan Hadoop dengan ukuran data 20GB, Objek Maksimum 30, dan Total 10 Core

Berdasarkan hasil grafik ( 5.12), dapat dilihat bahwa waktu eksekusi Spark menurun dan waktu eksekusi Hadoop meningkat ketika jumlah partisi diperbesar. Waktu eksekusi Hadoop menaik secara konsisten

ketika jumlah partisi diperbesar. Waktu eksekusi Spark menurun drastis pada awalnya ketika jumlah partisi ditingkatkan sampai titik tertentu dimana peningkatan jumlah partisi tidak memiliki dampak yang sangat drastis pada waktu eksekusi Spark. Tidak ada perbedaan yang jauh antara waktu eksekusi aplikasi Spark dengan standar deviasi maupun yang tidak. Aplikasi Spark memiliki waktu eksekusi yang lebih baik dibanding Hadoop pada jumlah partisi yang besar dan waktu eksekusi yang lebih buruk pada jumlah partisi yang kecil. Waktu eksekusi Spark pada partisi yang besar lebih cepat dibanding waktu eksekusi Hadoop terkecil. Aplikasi Spark lebih cepat dibanding Hadoop asalakan jumlah partisi diatur dengan benar.

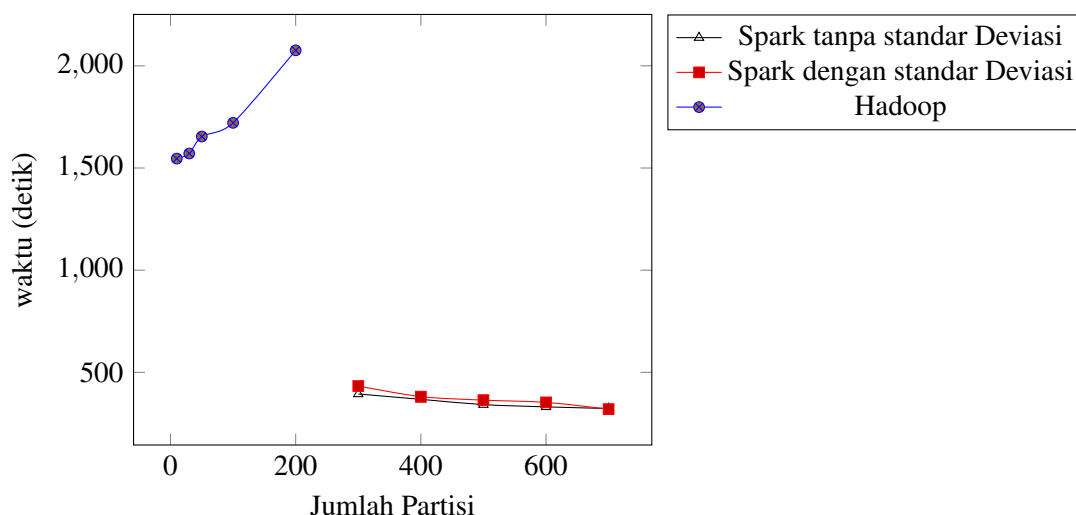
Pada percobaan ini akan dilihat waktu eksekusi Spark dan Hadoop berdasarkan jumlah partisi yang optimal. Percobaan ini akan menggunakan 1 komputer sebagai komputer *master* dan 10 komputer lainnya sebagai *worker* dengan setiap *worker* menggunakan 1 core. Ukuran data yang digunakan adalah 5 GB. Metode yang digunakan adalah metode *single linkage*, dengan nilai *cut-off distance* adalah 0,8 dan jumlah objek maksimum untuk setiap *dendrogram* adalah 50. Tabel ( 5.12) dan Tabel ( 5.13) berikut adalah hasil dari eksperimen:

Tabel 5.12: Percobaan Jumlah Partisi Hadoop dengan Ukuran Data 5 GB

Ukuran Data (GB)	Jumlah Partisi	Waktu Eksekusi Hadoop (detik)	Hasil Reduksi Hadoop (GB)
5	10	1546	1.5
5	30	1571	1.5
5	50	1654	1.5
5	100	1721	1.5
5	200	2076	1.5

Tabel 5.13: Percobaan Jumlah Partisi Spark dengan Ukuran Data 5 GB

Ukuran Data (GB)	Jumlah Partisi	Waktu Eksekusi Spark Tanpa standar Deviasi (detik)	Waktu Eksekusi Spark (detik)	Hasil Reduksi Spark standar Tanpa Deviasi (GB)	Hasil Reduksi Spark (GB)
5	300	394	433	1.5	1.9
5	400	368	381	1.5	1.9
5	500	342	364	1.5	1.9
5	600	331	353	1.5	1.9
5	700	323	320	1.5	1.9



Gambar 5.13: Hasil Percobaan Jumlah Partisi Spark dan Hadoop dengan Ukuran Data 5 GB, Objek Maksimum 50, dan Total 10 Core

Berdasarkan hasil grafik ( 5.13), dapat dilihat waktu Hadoop terus meningkat seiring meningkatnya jumlah partisi. Jumlah partisi yang dicoba pada Hadoop hanya mencapai 200 karena waktu eksekusi yang sudah berbeda jauh dibanding Spark dan untuk partisi yang lebih besar pasti diatas waktu eksekusi Hadoop dengan jumlah partisi sama dengan 200. Spark memiliki waktu eksekusi yang jauh lebih cepat dibanding Hadoop.

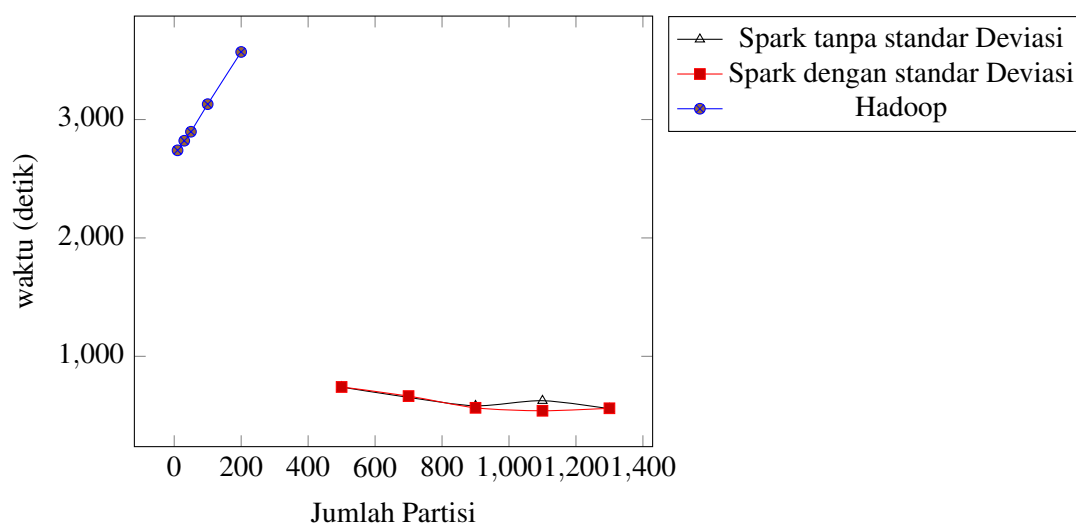
Pada percobaan ini akan dilihat waktu eksekusi Spark dan Hadoop berdasarkan jumlah partisi yang optimal. Percobaan ini akan menggunakan 1 komputer sebagai komputer *master* dan 10 komputer lainnya sebagai *worker* dengan setiap *worker* munggunakan 1 core. Ukuran data yang digunakan adalah 10 GB. Metode yang digunakan adalah metode *single linkage*, dengan nilai *cut-off distance* adalah 0,8 dan jumlah objek maksimum untuk setiap *dendrogram* adalah 50. Tabel ( 5.14) dan Tabel ( 5.15) berikut adalah hasil dari eksperimen:

Tabel 5.14: Percobaan Jumlah Partisi Hadoop dengan Ukuran Data 10 GB

Ukuran Data (GB)	Jumlah Partisi	Waktu Eksekusi Hadoop (detik)	Hasil Reduksi Hadoop (GB)
10	10	2740	2.7
10	30	2821	2.7
10	50	2897	2.7
10	100	3130	2.7
10	200	3571	2.7

Tabel 5.15: Percobaan Jumlah Partisi Spark dengan Ukuran Data 10 GB

Ukuran Data (GB)	Jumlah Partisi	Waktu Eksekusi Spark standar (detik)	Eksekusi Tanpa Deviasi	Waktu Eksekusi Spark (detik)	Hasil Spark standar (GB)	Reduksi Tanpa Deviasi	Hasil Reduksi Spark (GB)
10	500	740		741	2.7		3.3
10	700	653		664	2.7		3.3
10	900	582		565	2.7		3.3
10	1100	625		540	2.7		3.3
10	1300	557		561	2.7		3.3



Gambar 5.14: Hasil Percobaan Jumlah Partisi Spark dan Hadoop dengan Ukuran Data 10 GB, Objek Maksimum 50, dan Total 10 Core

Berdasarkan hasil grafik ( 5.14), dapat dilihat waktu Hadoop terus meningkat seiring meningkatnya jumlah partisi. Jumlah partisi yang dicoba pada Hadoop hanya mencapai 200 karena waktu eksekusi yang sudah berbeda jauh dibanding Spark dan untuk partisi yang lebih besar pasti diatas waktu eksekusi Hadoop dengan jumlah partisi sama dengan 200. Spark memiliki waktu eksekusi yang jauh lebih cepat dibanding Hadoop.

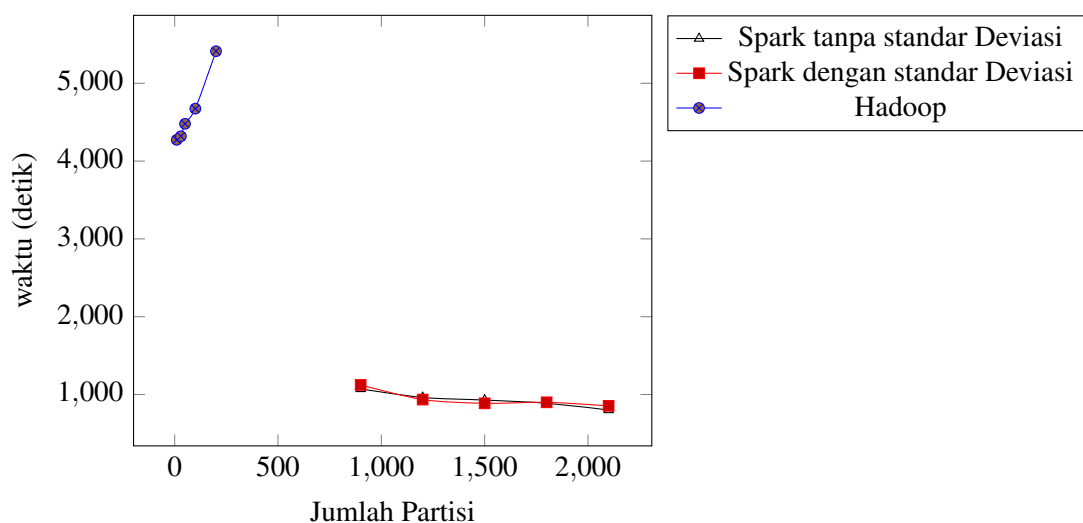
Pada percobaan ini akan dilihat waktu eksekusi Spark dan Hadoop berdasarkan jumlah partisi yang optimal. Percobaan ini akan menggunakan 1 komputer sebagai komputer *master* dan 10 komputer lainnya sebagai *worker* dengan setiap *worker* munggunakan 1 core. Ukuran data yang digunakan adalah 15 GB. Metode yang digunakan adalah metode *single linkage*, dengan nilai *cut-off distance* adalah 0,8 dan jumlah objek maksimum untuk setiap *dendrogram* adalah 50. Tabel ( 5.16) dan Tabel ( 5.17) berikut adalah hasil dari eksperimen:

Tabel 5.16: Percobaan Jumlah Partisi Hadoop dengan Ukuran Data 15 GB

Ukuran Data (GB)	Jumlah Partisi	Waktu Eksekusi Hadoop (detik)	Hasil Reduksi Hadoop (GB)
15	10	4273	4.2
15	30	4319	4.2
15	50	4479	4.2
15	100	4674	4.2
15	200	5412	4.2

Tabel 5.17: Percobaan Jumlah Partisi Spark dengan Ukuran Data 15 GB

Ukuran Data (GB)	Jumlah Partisi	Waktu Eksekusi Spark Tanpa standar Deviasi (detik)	Waktu Eksekusi Spark (detik)	Hasil Reduksi Spark standar (GB)	Hasil Reduksi Spark Tanpa Deviasi (GB)
15	900	1072	1123	4.2	5.2
15	1200	962	935	4.2	5.2
15	1500	929	887	4.2	5.2
15	1800	888	900	4.2	5.2
15	2100	801	854	4.2	5.2



Gambar 5.15: Hasil Percobaan Jumlah Partisi Spark dan Hadoop dengan Ukuran Data 15 GB, Objek Maksimum 50, dan Total 10 Core

Berdasarkan hasil grafik ( 5.15), dapat dilihat waktu Hadoop terus meningkat seiring meningkatnya jumlah partisi. Jumlah partisi yang dicoba pada Hadoop hanya mencapai 200 karena waktu eksekusi yang sudah berbeda jauh dibanding Spark dan untuk partisi yang lebih besar pasti diatas waktu eksekusi Hadoop dengan jumlah partisi sama dengan 200. Spark memiliki waktu eksekusi yang jauh lebih cepat dibanding Hadoop.

Pada percobaan ini akan dilihat waktu eksekusi Spark dan Hadoop berdasarkan jumlah partisi yang optimal. Percobaan ini akan menggunakan 1 komputer sebagai komputer *master* dan 10 komputer lainnya sebagai *worker* dengan setiap *worker* munggunakan 1 core. Ukuran data yang digunakan adalah 20 GB.



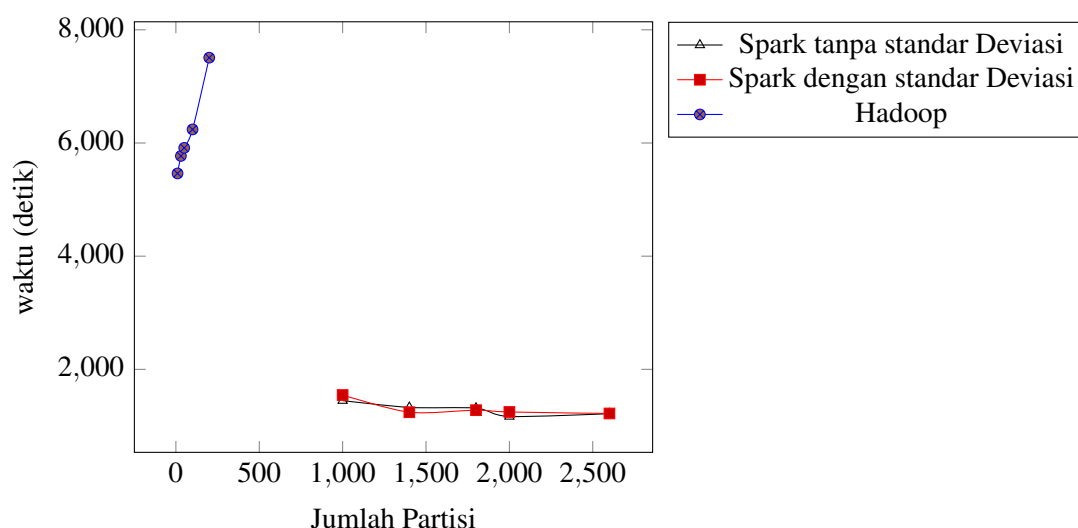
- 1 Metode yang digunakan adalah metode *single linkage*, dengan nilai *cut-off distance* adalah 0,8 dan jumlah
- 2 objek maksimum untuk setiap *dendrogram* adalah 50. Tabel ( 5.18) dan Tabel ( 5.19) berikut adalah hasil
- 3 dari eksperimen:

Tabel 5.18: Percobaan Jumlah Partisi Hadoop dengan Ukuran Data 20 GB

Ukuran Data (GB)	Jumlah Partisi	Waktu Eksekusi Hadoop (detik)	Hasil Reduksi Hadoop (GB)
20	10	5462	5.6
20	30	5771	5.6
20	50	5914	5.6
20	100	6240	5.6
20	200	7508	5.6

Tabel 5.19: Percobaan Jumlah Partisi Spark dengan Ukuran Data 20 GB

Ukuran Data (GB)	Jumlah Partisi	Waktu Eksekusi Spark standar (detik)	Waktu Eksekusi Spark Tanpa Deviasi (detik)	Hasil Reduksi Spark standar (GB)	Hasil Reduksi Spark Tanpa Deviasi (GB)
20	1000	1447	1546	5.6	6.9
20	1400	1327	1242	5.6	6.9
20	1800	1314	1278	5.6	6.9
20	2200	1167	1246	5.6	6.9
20	2600	1216	1220	5.6	6.9



Gambar 5.16: Hasil Percobaan Jumlah Partisi Spark dan Hadoop dengan Ukuran Data 20 GB, Objek Maksimum 50, dan Total 10 Core

Berdasarkan hasil grafik ( 5.16), dapat dilihat waktu Hadoop terus meningkat seiring meningkatnya jumlah partisi. Jumlah partisi yang dicoba pada Hadoop hanya mencapai 200 karena waktu eksekusi yang sudah berbeda jauh dibanding Spark dan untuk partisi yang lebih besar pasti diatas waktu eksekusi Hadoop dengan jumlah partisi sama dengan 200. Spark memiliki waktu eksekusi yang jauh lebih cepat dibanding Hadoop.

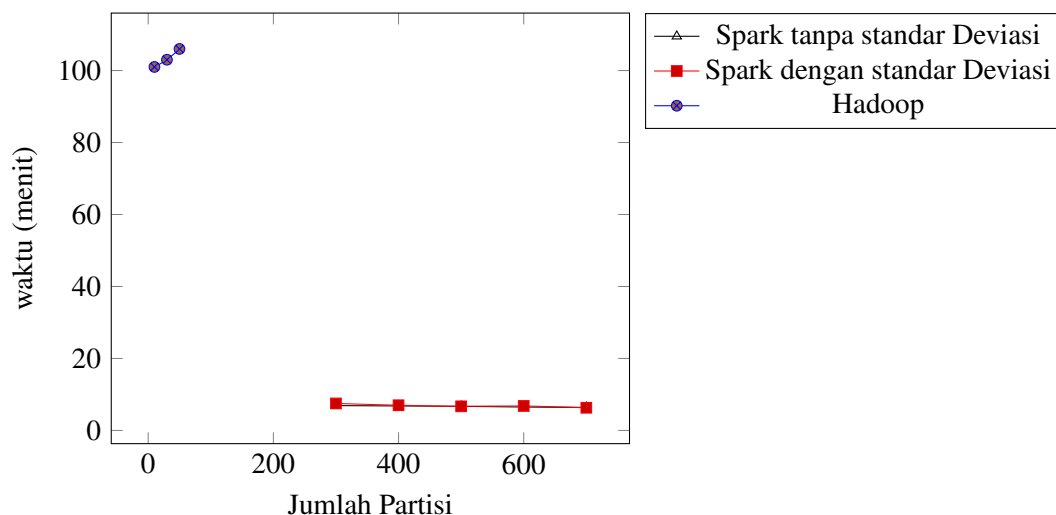
Pada percobaan ini akan dilihat waktu eksekusi Spark dan Hadoop berdasarkan jumlah partisi yang optimal. Percobaan ini akan menggunakan 1 komputer sebagai komputer *master* dan 10 komputer lainnya sebagai *worker* dengan setiap *worker* menggunakan 1 core. Ukuran data yang digunakan adalah 5 GB. Metode yang digunakan adalah metode *single linkage*, dengan nilai *cut-off distance* adalah 0,8 dan jumlah objek maksimum untuk setiap *dendrogram* adalah 100. Tabel ( 5.20) dan Tabel ( 5.21) berikut adalah hasil dari eksperimen:

Tabel 5.20: Percobaan Jumlah Partisi Hadoop dengan Ukuran Data 5 GB

Ukuran Data (GB)	Jumlah Partisi	Waktu Eksekusi Hadoop (menit)	Hasil Reduksi Hadoop (GB)
5	10	101	0.962
5	30	103	0.962
5	50	106	0.962

Tabel 5.21: Percobaan Jumlah Partisi Spark dengan Ukuran Data 5 GB

Ukuran Data (GB)	Jumlah Partisi	Waktu Eksekusi Spark standar (menit)	Waktu Eksekusi Spark Tanpa Deviasi (menit)	Hasil Reduksi Spark standar (GB)	Hasil Reduksi Spark Tanpa Deviasi (GB)
5	300	6.9	7.5	1	1.2
5	400	6.8	7.0	1	1.2
5	500	6.7	6.7	1	1.2
5	600	6.5	6.8	1	1.2
5	700	6.4	6.3	1	1.2



Gambar 5.17: Hasil Percobaan Jumlah Partisi Spark dan Hadoop dengan Ukuran Data 5 GB, Objek Maksimum 100, dan Total 10 Core

Berdasarkan hasil grafik ( 5.17), dapat dilihat waktu Hadoop terus meningkat seiring meningkatnya jumlah partisi. Jumlah partisi yang dicoba pada Hadoop hanya mencapai 200 karena waktu eksekusi yang sudah berbeda jauh dibanding Spark dan untuk partisi yang lebih besar pasti diatas waktu eksekusi Hadoop

dengan jumlah partisi sama dengan 200. Spark memiliki waktu eksekusi yang jauh lebih cepat dibanding Hadoop.

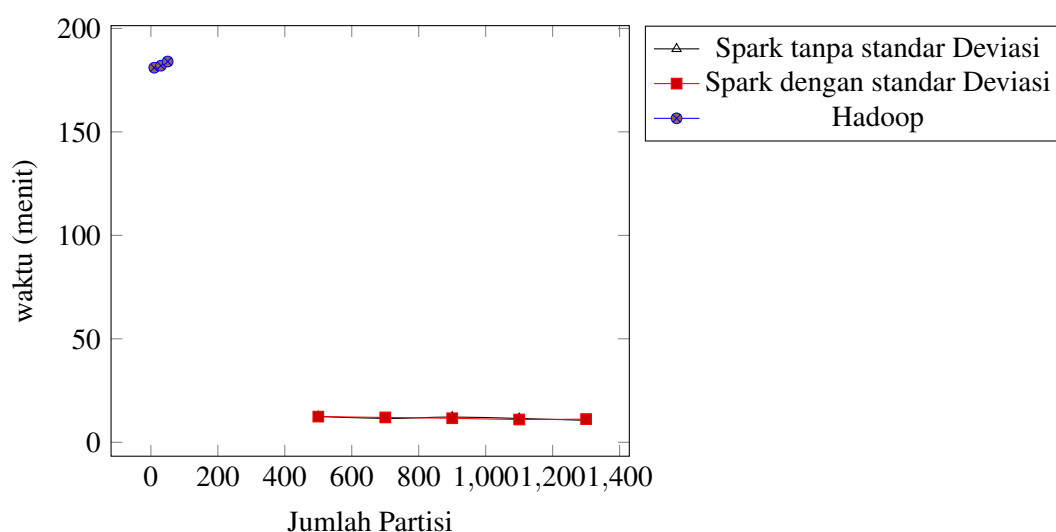
Pada percobaan ini akan dilihat waktu eksekusi Spark dan Hadoop berdasarkan jumlah partisi yang optimal. Percobaan ini akan menggunakan 1 komputer sebagai komputer *master* dan 10 komputer lainnya sebagai *worker* dengan setiap *worker* menggunakan 1 core. Ukuran data yang digunakan adalah 10 GB. Metode yang digunakan adalah metode *single linkage*, dengan nilai *cut-off distance* adalah 0,8 dan jumlah objek maksimum untuk setiap *dendrogram* adalah 100. Tabel ( 5.22) dan Tabel ( 5.23) berikut adalah hasil dari eksperimen:

Tabel 5.22: Percobaan Jumlah Partisi Hadoop dengan Ukuran Data 10 GB

Ukuran Data (GB)	Jumlah Partisi	Waktu Eksekusi Hadoop (menit)	Hasil Reduksi Hadoop (GB)
10	10	181	1.7
10	30	182	1.7
10	50	184	1.7

Tabel 5.23: Percobaan Jumlah Partisi Spark dengan Ukuran Data 10 GB

Ukuran Data (GB)	Jumlah Partisi	Waktu Eksekusi Spark standar (menit)	Waktu Eksekusi Spark Tanpa Deviasi (menit)	Hasil Reduksi Spark standar (GB)	Hasil Reduksi Spark Tanpa Deviasi (GB)
10	500	12.5	12.4	1.8	2.2
10	700	11.5	12.0	1.8	2.2
10	900	12.2	11.6	1.8	2.2
10	1100	11.5	11.0	1.8	2.2
10	1300	10.6	11.2	1.8	2.2



Gambar 5.18: Hasil Percobaan Jumlah Partisi Spark dan Hadoop dengan Ukuran Data 10 GB, Objek Maksimum 100, dan Total 10 Core

Berdasarkan hasil grafik ( 5.18), dapat dilihat waktu Hadoop terus meningkat seiring meningkatnya

1 jumlah partisi. Jumlah partisi yang dicoba pada Hadoop hanya mencapai 200 karena waktu eksekusi yang  
 2 sudah berbeda jauh dibanding Spark dan untuk partisi yang lebih besar pasti diatas waktu eksekusi Hadoop  
 3 dengan jumlah partisi sama dengan 200. Spark memiliki waktu eksekusi yang jauh lebih cepat dibanding  
 4 Hadoop.

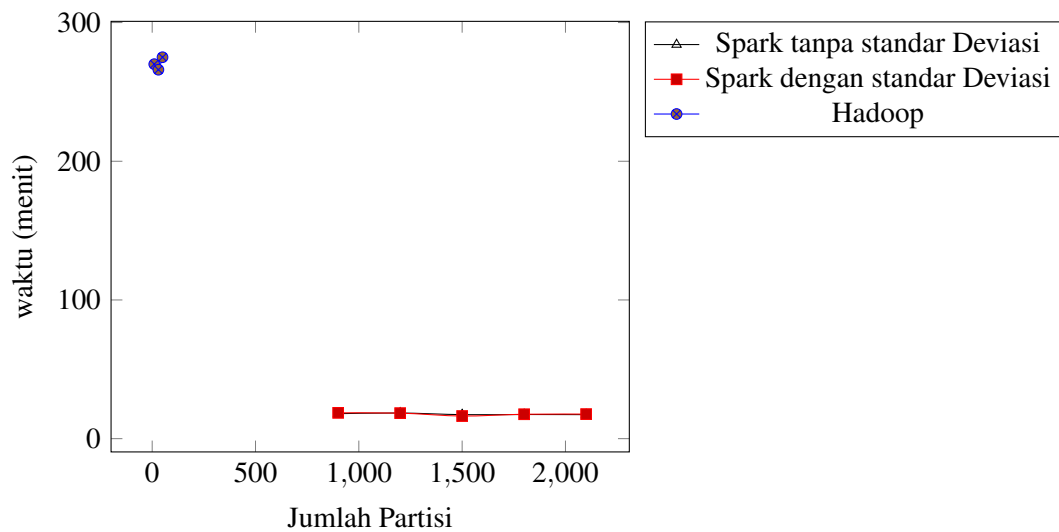
5  
 6 Pada percobaan ini akan dilihat waktu eksekusi Spark dan Hadoop berdasarkan jumlah partisi yang  
 7 optimal. Percobaan ini akan menggunakan 1 komputer sebagai komputer *master* dan 10 komputer lainnya  
 8 sebagai *worker* dengan setiap *worker* menggunakan 1 core. Ukuran data yang digunakan adalah 15 GB.  
 9 Metode yang digunakan adalah metode *single linkage*, dengan nilai *cut-off distance* adalah 0,8 dan jumlah  
 10 objek maksimum untuk setiap *dendrogram* adalah 100. Tabel ( 5.24) dan Tabel ( 5.25) berikut adalah hasil  
 11 dari eksperimen:

Tabel 5.24: Percobaan Jumlah Partisi Hadoop dengan Ukuran Data 15 GB

Ukuran Data (GB)	Jumlah Partisi	Waktu Eksekusi Hadoop (menit)	Hasil Reduksi Hadoop (GB)
15	10	270	2.6
15	30	266	2.6
15	50	275	2.6

Tabel 5.25: Percobaan Jumlah Partisi Spark dengan Ukuran Data 15 GB

Ukuran Data (GB)	Jumlah Partisi	Waktu Eksekusi Spark Tanpa standar Deviasi (menit)	Waktu Eksekusi Spark (menit)	Hasil Reduksi Spark standar Tanpa Deviasi (GB)	Hasil Reduksi Spark (GB)
15	900	18.0	18.6	2.8	3.4
15	1200	18.5	18.4	2.8	3.4
15	1500	17.3	16.3	2.8	3.4
15	1800	17.4	17.6	2.8	3.4
15	2100	17.3	17.7	2.8	3.4



Gambar 5.19: Hasil Percobaan Jumlah Partisi Spark dan Hadoop dengan Ukuran Data 15 GB, Objek Maksimum 100, dan Total 10 Core

Berdasarkan hasil grafik ( 5.19), dapat dilihat waktu Hadoop terus meningkat seiring meningkatnya jumlah partisi. Jumlah partisi yang dicoba pada Hadoop hanya mencapai 200 karena waktu eksekusi yang sudah berbeda jauh dibanding Spark dan untuk partisi yang lebih besar pasti diatas waktu eksekusi Hadoop dengan jumlah partisi sama dengan 200. Spark memiliki waktu eksekusi yang jauh lebih cepat dibanding Hadoop.

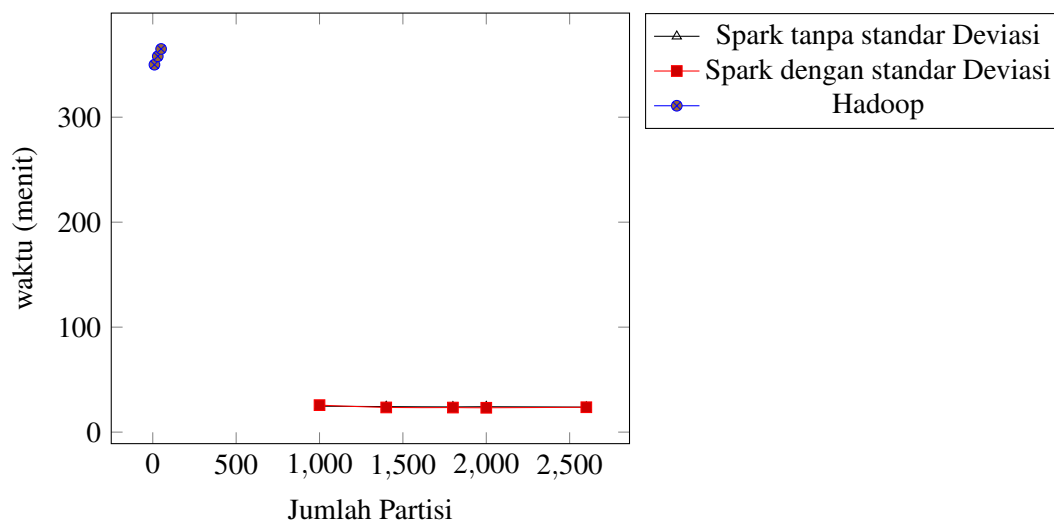
Pada percobaan ini akan dilihat waktu eksekusi Spark dan Hadoop berdasarkan jumlah partisi yang optimal. Percobaan ini akan menggunakan 1 komputer sebagai komputer *master* dan 10 komputer lainnya sebagai *worker* dengan setiap *worker* munggunakan 1 core. Ukuran data yang digunakan adalah 20 GB. Metode yang digunakan adalah metode *single linkage*, dengan nilai *cut-off distance* adalah 0,8 dan jumlah objek maksimum untuk setiap *dendrogram* adalah 100. Tabel ( 5.26) dan Tabel ( 5.27) berikut adalah hasil dari eksperimen:

Tabel 5.26: Percobaan Jumlah Partisi Hadoop dengan Ukuran Data 20 GB

Ukuran Data (GB)	Jumlah Partisi	Waktu Eksekusi Hadoop (menit)	Hasil Reduksi Hadoop (GB)
20	10	350	3.5
20	30	358	3.5
20	50	365	3.5

Tabel 5.27: Percobaan Jumlah Partisi Spark dengan Ukuran Data 20 GB

Ukuran Data (GB)	Jumlah Partisi	Waktu Eksekusi Spark standar (menit)	Waktu Eksekusi Spark Tanpa Deviasi	Waktu Eksekusi Spark (menit)	Hasil Spark standar (GB)	Hasil Spark Tanpa Deviasi	Hasil Spark (GB)
20	1000	24.7		25.7	3.7		4.5
20	1400	24.3		23.5	3.7		4.5
20	1800	24.0		23.4	3.7		4.5
20	2200	24.2		23.2	3.7		4.5
20	2600	23.8		23.7	3.7		4.5



Gambar 5.20: Hasil Percobaan Jumlah Partisi Spark dan Hadoop dengan Ukuran Data 20 GB, Objek Maksimum 100, dan Total 10 Core

Berdasarkan hasil grafik ( 5.20), dapat dilihat waktu Hadoop terus meningkat seiring meningkatnya jumlah partisi. Jumlah partisi yang dicoba pada Hadoop hanya mencapai 200 karena waktu eksekusi yang sudah berbeda jauh dibanding Spark dan untuk partisi yang lebih besar pasti diatas waktu eksekusi Hadoop dengan jumlah partisi sama dengan 200. Spark memiliki waktu eksekusi yang jauh lebih cepat dibanding Hadoop.

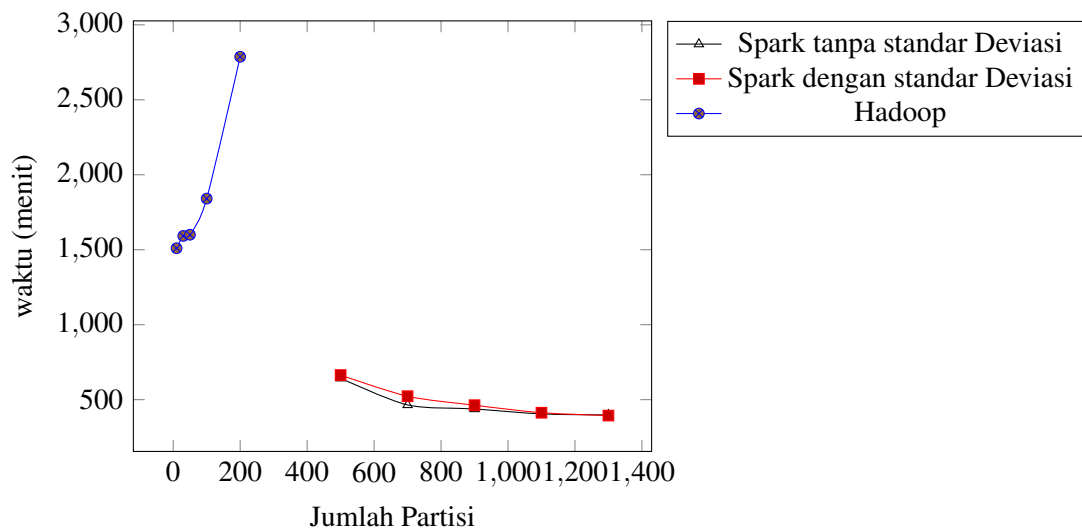
Pada percobaan ini akan dilihat waktu eksekusi Spark dan Hadoop berdasarkan jumlah partisi yang optimal. Percobaan ini akan menggunakan 1 komputer sebagai komputer *master* dan 10 komputer lainnya sebagai *worker* dengan setiap *worker* munggunakan 1 core. Ukuran data yang digunakan adalah 10 GB. Metode yang digunakan adalah metode *complete linkage*, dengan nilai *cut-off distance* adalah 0,8 dan jumlah objek maksimum untuk setiap *dendrogram* adalah 30. Tabel ( 5.28) dan Tabel ( 5.29) berikut adalah hasil dari eksperimen:

Tabel 5.28: Percobaan Jumlah Partisi Hadoop dengan Ukuran Data 10 GB

Ukuran Data (GB)	Jumlah Partisi	Waktu Eksekusi Hadoop (detik)	Hasil Reduksi Hadoop (GB)
10	10	1510	2.4
10	30	1592	2.4
10	50	1600	2.4
10	100	1841	2.4
10	200	2787	2.4

Tabel 5.29: Percobaan Jumlah Partisi Spark dengan Ukuran Data 10 GB

Ukuran Data (GB)	Jumlah Partisi	Waktu Eksekusi Spark Tanpa standar Deviasi (detik)	Waktu Eksekusi Spark (detik)	Hasil Reduksi Spark standar (GB)	Hasil Reduksi Spark Tanpa Deviasi (GB)
10	500	643	664	2.3	3.0
10	700	465	523	2.3	3.0
10	900	438	463	2.3	3.0
10	1100	405	413	2.3	3.0
10	1300	398	394	2.3	3.0



Gambar 5.21: Hasil Percobaan Jumlah Partisi Spark dan Hadoop dengan Ukuran Data 20 GB, Objek Maksimum 30, dan Total 10 Core

Berdasarkan hasil grafik ( 5.21), dapat dilihat waktu Hadoop terus meningkat seiring meningkatnya jumlah partisi. Jumlah partisi yang dicoba pada Hadoop hanya mencapai 200 karena waktu eksekusi yang sudah berbeda jauh dibanding Spark dan untuk partisi yang lebih besar pasti diatas waktu eksekusi Hadoop dengan jumlah partisi sama dengan 200. Spark memiliki waktu eksekusi yang jauh lebih cepat dibanding Hadoop.

Pada percobaan ini akan dilihat waktu eksekusi Spark dan Hadoop berdasarkan jumlah partisi yang optimal. Percobaan ini akan menggunakan 1 komputer sebagai komputer *master* dan 10 komputer lainnya sebagai *worker* dengan setiap *worker* menggunakan 1 core. Ukuran data yang digunakan adalah 10 GB.

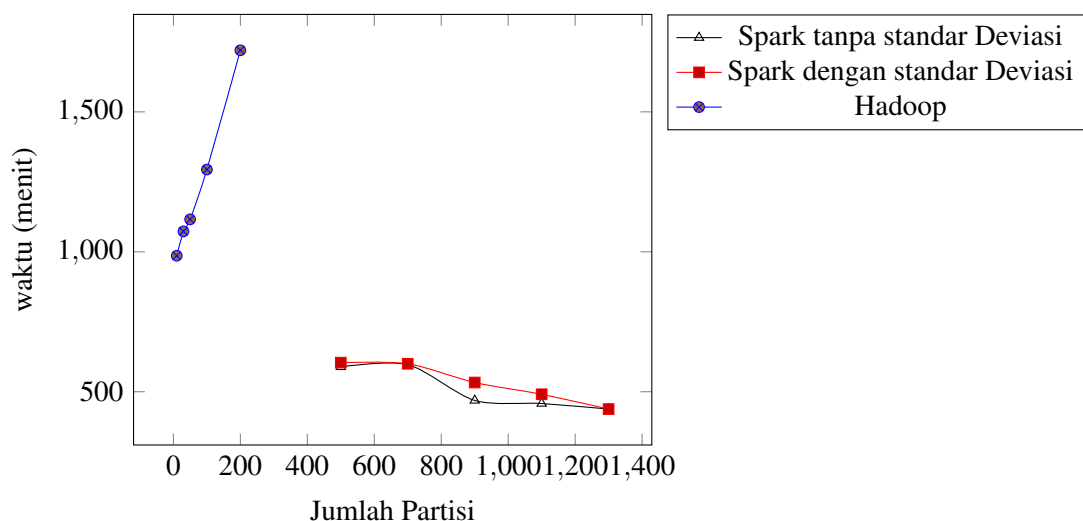
- 1 Metode yang digunakan adalah metode *centroid linkage*, dengan nilai *cut-off distance* adalah 0,8 dan jumlah
- 2 objek maksimum untuk setiap *dendrogram* adalah 30. Tabel ( 5.30) dan Tabel ( 5.31) berikut adalah hasil
- 3 dari eksperimen:

Tabel 5.30: Percobaan Jumlah Partisi Hadoop dengan Ukuran Data 10 GB

Ukuran Data (GB)	Jumlah Partisi	Waktu Eksekusi Hadoop (detik)	Hasil Reduksi Hadoop (GB)
10	10	986	2.7
10	30	1073	2.7
10	50	1116	2.7
10	100	1294	2.7
10	200	1720	2.7

Tabel 5.31: Percobaan Jumlah Partisi Spark dengan Ukuran Data 10 GB

Ukuran Data (GB)	Jumlah Partisi	Waktu Eksekusi Spark standar (detik)	Waktu Eksekusi Spark Tanpa Deviasi (detik)	Hasil Reduksi Spark standar (GB)	Hasil Reduksi Spark Tanpa Deviasi (GB)
10	500	589	604	2.8	3.7
10	700	597	600	2.8	3.7
10	900	469	533	2.8	3.7
10	1100	458	491	2.8	3.7
10	1300	438	438	2.8	3.7



Gambar 5.22: Hasil Percobaan Jumlah Partisi Spark dan Hadoop dengan Ukuran Data 20 GB, Objek Maksimum 30, dan Total 10 Core

- 5 Berdasarkan hasil grafik ( 5.22), dapat dilihat waktu Hadoop terus meningkat seiring meningkatnya
- 6 jumlah partisi. Jumlah partisi yang dicoba pada Hadoop hanya mencapai 200 karena waktu eksekusi yang
- 7 sudah berbeda jauh dibanding Spark dan untuk partisi yang lebih besar pasti diatas waktu eksekusi Hadoop
- 8 dengan jumlah partisi sama dengan 200. Spark memiliki waktu eksekusi yang jauh lebih cepat dibanding
- 9 Hadoop.



1 Berdasarkan eksperimen-eksperimen diatas, dapat disimpulkan bahwa Spark dapat bekerja lebih cepat  
2 dibanding Hadoop pada partisi yang besar. Sebaliknya, waktu eksekusi Hadoop meningkat ketika jumlah  
3 partisi ditingkatkan. Waktu eksekusi Hadoop terbaik dapat masih lebih tinggi dibanding waktu eksekusi  
4 terbaik Spark. Meningkatnya waktu eksekusi Hadoop disebabkan oleh *shuffling* dan *sorting*. Semakin tinggi  
5 jumlah partisi, semakin banyak yang di-*shuffle* dan di-*sorting*. Berbeda dengan Hadoop, waktu eksekusi  
6 Spark menurun ketika jumlah partisi ditingkatkan. Dengan meningkatkan jumlah partisi pada Spark, data  
7 akan lebih terdistribusi. Hal ini akan mengurangi waktu yang dibutuhkan untuk mengirim data dari satu  
8 komputer ke komputer lain dan mengurangi waktu komunikasi.



## BAB 6

### KESIMPULAN DAN SARAN

Bab ini berisi kesimpulan dari awal hingga akhir penelitian beserta saran untuk penelitian selanjutnya.

#### 6.1 Kesimpulan

Kesimpulan yang dapat ditarik dari awal penelitian ini sampai selesai adalah sebagai berikut:

- Pada penelitian ini, telah dipelajari algoritma *Hierarchical Agglomerative Clustering*.
- Pada penelitian ini, telah diimplementasikan algoritma *Hierarchical Agglomerative Clustering* pada lingkungan Spark dengan menggunakan *transformation* dan *actions*. Fungsi *map()*, *groupByKey()*, *flatMap()* dapat digunakan untuk menggantikan fungsi *map()* dan *reduce()* pada MapReduce.
- Pada penelitian ini, telah dilakukan eksperimen perbandingan performa antara perangkat lunak Spark dan Hadoop. Dari hasil pengujian dapat disimpulkan bahwa perangkat lunak Spark memiliki performa yang lebih baik asalkan diatur dan dikonfigurasi dengan benar. Waktu eksekusi Spark lebih cepat dibanding Hadoop karena Spark menyimpan data pada memori, sebaliknya Hadoop banyak melakukan proses I/O kepada disk yang membuat Hadoop lambat. Proses *shuffling* dan *sorting* menghambat dan meningkatkan waktu eksekusi Hadoop ketika jumlah partisi ditingkatkan. Waktu eksekusi Hadoop akan meningkat seiring meningkatnya jumlah partisi. Sebaliknya, waktu eksekusi Spark menurun ketika jumlah partisi ditingkatkan. Dengan meningkatkan jumlah partisi pada Spark, data akan lebih terdistribusi. Hal ini akan mengurangi waktu yang dibutuhkan untuk mengirim data dari satu komputer ke komputer lain dan mengurangi waktu komunikasi.
- Pada penelitian ini, telah dibangun perangkat lunak untuk melihat hasil reduksi data dan menjalankan proses reduksi data.

#### 6.2 Saran

Saran untuk penelitian selanjutnya adalah sebagai berikut:

- Pada penelitian ini, Spark dijalankan pada Hadoop *YARN*. Oleh karena itu, penulis berharap agar penelitian selanjutnya dapat menguji performa perangkat lunak pada Spark *cluster* atau *cluster* lainnya.
- Pada penelitian ini, pengujian yang dilakukan masih terbatas dengan 10 *worker* dan ukuran data sampai 20GB. Untuk penelitian selanjutnya, penulis berharap agar pengujian yang dilakukan dapat menggunakan jumlah *worker* dan data yang lebih besar.



## DAFTAR REFERENSI

- [1] Moertini, V. S., Suarjana, G. W., Venica, L., dan Karya, G. (2018) Big data reduction technique using parallel hierarchical agglomerative clustering. *IAENG International Journal of Computer Science*, **45**, 188 – 205.
- [2] Ishwarappa dan J, A. (2015) A brief introduction on big data 5vs characteristics and hadoop technology. *Procedia Computer Science*, **48**, 319 – 324.
- [3] Jain, A. K. dan Dubes, R. C. (1988) *Algorithms for Clustering Data*. Pearson College Div, New Jersey.
- [4] Holmes, A. (2012) *Hadoop in Practice*. Manning, New York.
- [5] White, T. (2015) *Hadoop The Definitive Guide*, 4th edition. O'Reilly Media, Sebastopol.
- [6] Lam, C. (2010) *Hadoop in Action*. Manning Publications, New York.
- [7] Karau, H., Konwinski, A., Wndell, P., dan Zaharia, M. (2015) *Learning Spark*, 1th edition. O'Reilly Media, Sebastopol.



# LAMPIRAN A

## KODE PROGRAM

Listing A.1: Main.scala

```
1 package main.scala
2 import org.apache.spark.{SparkConf, SparkContext}
3
4 object Main {
5   def main(args: Array[String]): Unit = {
6     val master = "yarn-cluster"
7     val input = args(0)
8     val output = args(1)
9     val numPar = args(2).toInt
10    val maxObj = args(3).toInt
11    val distType = args(4).toInt
12    val cutOffDist = args(5).toDouble
13    val conf = new SparkConf()
14    conf.setMaster(master)
15    conf.setAppName("Reduce_Data_Spark")
16    val sc = new SparkContext(conf)
17    val dataReducer = new DataReducer(sc, numPar, maxObj, distType, cutOffDist, input, output)
18    dataReducer.reduceData()
19  }
20 }
```

Listing A.2: DataReducer.scala

```
1 package main.scala
2
3 import org.apache.spark.{SparkContext}
4 import org.apache.spark.rdd.RDD
5 import scala.collection.mutable.{ListBuffer}
6
7 class DataReducer(sc: SparkContext, numPar: Int, maxObj: Int, distanceType: Int, cutOffDistance: Double, inputPath: String, outputPath: String) extends Serializable {
8
9   def reduceData(): Unit = {
10     val broadCastMaxObj = sc.broadcast(maxObj)
11     val broadCastDistanceType = sc.broadcast(distanceType)
12     val broadCastCutOffDist = sc.broadcast(cutOffDistance)
13     val result = mapData.groupByKey(numPar).map(record => {
14       val patterns: ListBuffer[Pattern] = new ListBuffer[Pattern]()
15       var i: Int = 0
16       var isProcessed: Boolean = false
17       var objectList: ListBuffer[Node] = new ListBuffer[Node]()
18       record._2.foreach(record => {
19         isProcessed = false
20         i += 1
21         objectList += record
22         if (i == broadCastMaxObj.value) {
23           val dendrogram: Dendrogram = new Dendrogram(objectList, broadCastDistanceType.value)
24           dendrogram.generateDendrogram()
25           val cluster = new Cluster(dendrogram.getDendrogram(), broadCastCutOffDist.value)
26           patterns = patterns ++ cluster.computePatern()
27           isProcessed = true
28           i = 0
29           objectList.clear()
30         }
31       })
32       if (isProcessed == false) {
33         val dendrogram: Dendrogram = new Dendrogram(objectList, broadCastDistanceType.value)
34         dendrogram.generateDendrogram()
35         val cluster = new Cluster(dendrogram.getDendrogram(), broadCastCutOffDist.value)
36         patterns = patterns ++ cluster.computePatern()
37       }
38       patterns.toIterator
39     })
40
41     val parseResult = result.flatMap(patterns => {
42       patterns.map(pattern => pattern)
43     })
44
45     val finalResult = parseResult.map(pattern => { pattern.getObjCount() + "\n" +
46       pattern.getMinArr().mkString(",") + "\n" +
47       pattern.getMaxArr().mkString(",") + "\n" +
48       pattern.getAvgArr().mkString(",") + "\n" +
49       pattern.getSDArr().mkString(",")
50     }).saveAsTextFile(outputPath)
```

```

51      broadcastMaxObj.destroy()
52      broadcastDistanceType.destroy()
53      broadcastCutOffDist.destroy()
54      sc.stop()
55  }
56
57  }
58
59  private def loadData():RDD[String] = {
60      sc.textFile(inputPath)
61  }
62
63  private def mapData():RDD[(Int,Node)] = {
64      val broadcastNpar = sc.broadcast(numPar)
65      val result = loadData().map(lines => {
66          val node = new Node()
67          node.setData(lines.split(",").map(_.toDouble))
68          val key = Random.nextInt(broadcastNpar.value)
69          (key,node)
70      })
71      result
72  }
73  }

```

Listing A.3: Dendrogram.scala

```

1  package main.scala
2
3  import scala.collection.mutable.{ArrayBuffer, ListBuffer}
4
5  class Dendrogram(nodeList:ListBuffer[Node], distType:Int) extends Serializable {
6      private var dendrogram = new ArrayBuffer[Node]()
7      private var nodeListCluster = new ArrayBuffer[ListBuffer[Node]]()
8      private var distanceMatrix = new ArrayBuffer[ArrayBuffer[Double]]()
9
10     def getDendrogram(): Node = {
11         dendrogram(0)
12     }
13
14     def generateDendrogram(): Unit = {
15         var i = 0
16         nodeList.foreach(node => {
17             dendrogram += node
18             nodeListCluster += new ListBuffer[Node]
19             nodeListCluster(i) += node
20             distanceMatrix += new ArrayBuffer[Double]()
21             i+=1
22         })
23         i = 1
24         var x = 0
25         for(i <- 1 until distanceMatrix.length){
26             for(x <- 0 until i){
27                 distanceMatrix(i) += findMinimumDistance(nodeListCluster(i),nodeListCluster(x))
28             }
29         }
30
31         while(dendrogram.length !=1){
32             var x = 1
33             var y = 0
34             var result = Double.MaxValue
35             var coordinateX = 0
36             var coordinateY = 0
37             var temp = 0.0
38             for(x <- 1 until distanceMatrix.length){
39                 for(y <- 0 until x){
40                     temp = distanceMatrix(x)(y)
41                     if(temp < result){
42                         result = temp
43                         coordinateX = x
44                         coordinateY = y
45                     }
46                 }
47             }
48             formClusterBetweenNearestNeighbour(coordinateX,coordinateY)
49             recalculateMatrix(coordinateX,coordinateY)
50         }
51     }
52
53     private def formClusterBetweenNearestNeighbour(x:Int,y:Int): Unit = {
54         nodeListCluster(y) = nodeListCluster(y) ++ nodeListCluster(x)
55         nodeListCluster.remove(x)
56         val cluster = new Node()
57         cluster.setDistance(distanceMatrix(x)(y))
58         cluster.setLeftNode(dendrogram(y))
59         cluster.setRightNode(dendrogram(x))
60         dendrogram(y) = cluster
61         dendrogram.remove(x)
62     }
63
64     private def recalculateMatrix(x:Int,y:Int): Unit = {
65         distanceMatrix.remove(x)
66         for(i <- x+1 until distanceMatrix.length){
67             distanceMatrix(i).remove(x)
68         }
69         for(i <- y+1 until distanceMatrix.length){
70             distanceMatrix(i)(y) = findMinimumDistance(nodeListCluster(i), nodeListCluster(y))
71         }
72     }

```



```

73 | }
74 |
75 | private def findMinimumDistance(firstList:ListBuffer[Node],secondList:ListBuffer[Node]): Double = {
76 |   if(distType == 0) calculateSingleLinkage(firstList,secondList)
77 |   else if (distType == 1) calculateCompleteLinkage(firstList, secondList)
78 |   else calculateCentroidLinkage(firstList,secondList)
79 | }
80 |
81 | private def calculateCentroidLinkage(firstList:ListBuffer[Node], secondList:ListBuffer[Node]): Double = {
82 |   val length = firstList(0).getData().length
83 |   val firstArr = new Array[Double](length)
84 |   val secondArr = new Array[Double](length)
85 |   var i = 0
86 |   var max = firstList.length
87 |   if(secondList.length > max) max = secondList.length
88 |   while(i < max){
89 |     if(i < firstList.length ){
90 |       var index = 0;
91 |       firstList(i).getData().foreach( data => {
92 |         firstArr(index) += data
93 |         index+=1
94 |       })
95 |     }
96 |     if(i < secondList.length){
97 |       var index = 0;
98 |       secondList(i).getData().foreach( data => {
99 |         secondArr(index) += data
100 |         index+=1
101 |       })
102 |     }
103 |     i+=1
104 |   }
105 |   i=0
106 |   while(i<firstArr.length){
107 |     firstArr(i) /= firstList.length
108 |     secondArr(i) /= secondList.length
109 |     i+=1
110 |   }
111 |   calculateDistance(firstArr,secondArr)
112 | }
113 |
114 | private def calculateSingleLinkage(firstList:ListBuffer[Node], secondList:ListBuffer[Node]): Double = {
115 |   var min:Double = Double.MaxValue
116 |   var result:Double = 0
117 |   firstList.foreach( nodeA => {
118 |     secondList.foreach( nodeB => {
119 |       result = calculateDistance(nodeA.getData(), nodeB.getData())
120 |       if(result < min) min = result
121 |     })
122 |   })
123 |   min
124 | }
125 |
126 | private def calculateCompleteLinkage(firstList:ListBuffer[Node], secondList:ListBuffer[Node]): Double = {
127 |   var max:Double = Double.MinValue
128 |   var result:Double = 0
129 |   firstList.foreach( nodeA => {
130 |     secondList.foreach( nodeB => {
131 |       result = calculateDistance(nodeA.getData(), nodeB.getData())
132 |       if(result > max) max = result
133 |     })
134 |   })
135 |   max
136 | }
137 |
138 | private def calculateDistance(firstArr:Array[Double], secondArr:Array[Double]): Double = {
139 |   val n = firstArr.length-1
140 |   var total:Double = 0
141 |   for(i <- 0 to n){
142 |     total +=Math.pow(firstArr(i)-secondArr(i),2)
143 |   }
144 |   Math.sqrt(total)
145 | }
146 | }

```

Listing A.4: Cluster.scala

```

1 | package main.scala
2 |
3 | import scala.collection.mutable.{ArrayBuffer, ListBuffer}
4 |
5 | class Cluster(dendrogram:Node, cutOffDistance:Double) extends Serializable {
6 |   private val clusters:ListBuffer[Node] = new ListBuffer[Node]()
7 |
8 |   private def formClusterFromDendrogram(): Unit = {
9 |     val bfs:ListBuffer[Node] = new ListBuffer[Node]
10 |     bfs+=dendrogram
11 |     val distance = cutOffDistance * dendrogram.getDistance()
12 |     while(bfs.length!=0){
13 |       var node = bfs.remove(0)
14 |       if(node.getDistance() <= distance){
15 |         clusters+=node
16 |       } else {
17 |         var left = node.getLeftNode()
18 |         var right = node.getRightNode()
19 |         if(left!=null){
20 |           bfs+=left
21 |         }

```

```

22         if(right!=null){
23             bfs+=right
24         }
25     }
26 }
27
28
29 def computePatern(): ListBuffer[Patern] = {
30     formClusterFromDendrogram()
31     val paterns:ListBuffer[Patern] = new ListBuffer[Patern]()
32     clusters.foreach( cluster => {
33         paterns += processCluster(cluster)
34     })
35     paterns
36 }
37
38 private def processCluster(cluster: Node): Patern ={
39     val bfs:ListBuffer[Node] = new ListBuffer[Node]()
40     val min:ArrayBuffer[Double] = new ArrayBuffer[Double]()
41     val max:ArrayBuffer[Double] = new ArrayBuffer[Double]()
42     val avg:ArrayBuffer[Double] = new ArrayBuffer[Double]()
43     val SD:ArrayBuffer[Double] = new ArrayBuffer[Double]()
44     bfs+=cluster
45     var count = 0
46     var i=0
47     while(bfs.length!=0){
48         val node = bfs.remove(0)
49         val data = node.getData()
50         if(data!=null){
51             if(min.length==0){
52                 data.foreach(value => {
53                     min+=value
54                     max+=value
55                     avg+=value
56                 })
57             } else {
58                 i=0
59                 data.foreach(value => {
60                     if(value < min(i)) min(i) = value
61                     if(value > max(i)) max(i) = value
62                     avg(i) += value
63                     i+=1
64                 })
65             }
66             count+=1
67         } else {
68             val leftNode = node.getLeftNode()
69             val rightNode = node.getRightNode()
70             if(leftNode!=null){
71                 bfs+=leftNode
72             }
73             if(rightNode!=null){
74                 bfs+=rightNode
75             }
76         }
77     }
78     i =0;
79     avg.foreach( value => {
80         avg(i) /= count
81         i+=1
82     })
83     bfs+=cluster
84     while(bfs.length!=0){
85         val node = bfs.remove(0)
86         val data = node.getData()
87         if(data!=null){
88             if(SD.length==0){
89                 i=0
90                 data.foreach(value => {
91                     //println("TEST SD")
92                     //println(value+" "+avg(i))
93                     SD += Math.pow((value - avg(i)),2)
94                     //println(SD(0))
95                     i+=1
96                 })
97             } else {
98                 i=0
99                 data.foreach(value => {
100                     SD(i) += Math.pow((value - avg(i)),2)
101                     i+=1
102                 })
103             }
104         } else {
105             val leftNode = node.getLeftNode()
106             val rightNode = node.getRightNode()
107             if(leftNode!=null){
108                 bfs+=leftNode
109             }
110             if(rightNode!=null){
111                 bfs+=rightNode
112             }
113         }
114     }
115     i =0;
116     SD.foreach( value => {
117         if(count == 1){
118             SD(i) = 0
119         } else {
120             SD(i) = Math.sqrt((SD(i) / (count - 1)));

```

```

121|         i += 1
122|     }
123| })
124| new Patern(max.toArray,min.toArray,avg.toArray,SD.toArray,count)
125| }
126| }

```

Listing A.5: Node.scala

```

1| package main.scala
2|
3| class Node() extends Serializable {
4|     private var data:Array[Double] = null
5|     private var distance:Double = -1
6|     private var rightNode:Node = null
7|     private var leftNode:Node = null
8|
9|     def setData(data: Array[Double]): Unit = {
10|         this.data = data
11|     }
12|
13|     def setDistance(distance: Double): Unit = {
14|         this.distance = distance
15|     }
16|
17|     def setRightNode(node:Node): Unit = {
18|         this.rightNode = node
19|     }
20|
21|     def setLeftNode(node:Node): Unit = {
22|         this.leftNode = node
23|     }
24|
25|     def getData(): Array[Double] = {
26|         this.data
27|     }
28|
29|     def getDistance():Double = {
30|         this.distance
31|     }
32|
33|     def getRightNode(): Node = {
34|         this.rightNode
35|     }
36|
37|     def getLeftNode(): Node = {
38|         this.leftNode
39|     }
40| }

```

Listing A.6: Node.scala

```

1| package main.scala
2|
3| class Patern(max:Array[Double], min:Array[Double], avg:Array[Double], SD:Array[Double], objCount:Int) extends Serializable {
4|
5|     def getMaxArr(): Array[Double] = {
6|         max
7|     }
8|
9|     def getMinArr(): Array[Double] = {
10|         min
11|     }
12|
13|     def getAvgArr(): Array[Double] = {
14|         avg
15|     }
16|
17|     def getSDArr(): Array[Double] = {
18|         SD
19|     }
20|
21|     def getObjCount(): Int = {
22|         objCount
23|     }
24| }

```



## LAMPIRAN B

### KODE PROGRAM UNTUK ANTARMUKA

Listing B.1: index.php

```
1 <!DOCTYPE html>
2 <html>
3 <head>
4   <?php include 'head.php'; ?>
5   <title>Spark Reduce Data App UI</title>
6   <style type="text/css">
7     .input-field label {
8       color: #212121;
9     }
10    /* label focus color */
11    .input-field input[type=text]:focus + label {
12      color: #212121;
13    }
14    /* label underline focus color */
15    .input-field input[type=text]:focus {
16      border-bottom: 1px solid #212121;
17      box-shadow: 0 1px 0 0 #212121;
18    }
19    /* icon prefix focus color */
20    .input-field .prefix.active {
21      color: #212121;
22    }
23  </style>
24 </head>
25
26 <body>
27   <?php include 'nav.php' ?>
28   <div class="row">
29     <div class="col_m4_offset-m4">
30       <div class="row">
31         <div class="col_m12" style="margin-top:_40px;">
32           <div class="card_teal_lighten-5">
33             <div class="card-content_black-text">
34               <span class="card-title">Submit Spark Application</span>
35             <div class="row">
36               <?php include 'form.php'; ?>
37             </div>
38           </div>
39         </div>
40       </div>
41     </div>
42   </div>
43 </body>
44 </html>
45 <script type="text/javascript">
46   $(document).ready(function(){
47
48     $('select').formSelect();
49
50     $( "#spark-form" ).submit(function( event ) {
51       window.open( 'master:8080/cluster', '_blank' );
52     });
53   });
54 </script>
```

Listing B.2: head.php

```
1 <link rel="stylesheet" href="https://cdnjs.cloudflare.com/ajax/libs/materialize/1.0.0/css/materialize.min.css">
2 <script src="https://ajax.googleapis.com/ajax/libs/jquery/3.4.0/jquery.min.js"></script>
3 <script src="https://cdnjs.cloudflare.com/ajax/libs/materialize/1.0.0/js/materialize.min.js"></script>
```

Listing B.3: data.php

```
1 <!DOCTYPE html>
2 <html>
3 <head>
4   <?php include 'head.php' ?>
5   <title>Spark Reduce Data App UI</title>
6   <style type="text/css">
7   </style>
8 </head>
9
```

```

10 <body>
11 <?php include 'nav.php' ?>
12 <div class="row">
13 <div class="col">
14 <div class="row">
15 <div class="col_m12" style="margin-top:10px;">
16
17 <?php
18 $temp = "00000";
19 $sub = substr($temp, strlen($_GET['part']));
20 //echo $sub;
21 //echo $_GET['part'];
22 $output = shell_exec('cd_/home/miebakso/hadoop-2.7.3_&&_/bin/hadoop_fs_cat_' . $_GET['path'] . '/part-' . $sub .
    $_GET['part']);
23 $arr = explode("\n", $output);
24 $i = 1;
25 $len = count(explode(',', $arr[1]));
26 foreach ($arr as $value) {
27     if($i==1){
28         echo '
29         <div class="card_teal_lighten-4">
30         <div class="card-content_black-text">
31         <table>
32         <thead>
33         <tr>
34         <th>Total_Obj_ ' . $value . ' </th>';
35         $x=1;
36         while($x<=$len){
37             echo ' <th>attribute- ' . $x . ' </th>';
38             $x++;
39         }
40
41         echo '
42         </tr>
43         </thead>
44
45         <tbody>
46         '
47         ;
48
49     } else if($i==2){
50         $arr2 = explode(",", $value);
51         echo ' <tr><td>Minimum</td>';
52         foreach ($arr2 as $val) {
53             echo ' <td>' . $val . ' </td>';
54         }
55         echo ' </tr>';
56     } else if($i==3){
57         $arr2 = explode(",", $value);
58         echo ' <tr><td>Maximum</td>';
59         foreach ($arr2 as $val) {
60             echo ' <td>' . $val . ' </td>';
61         }
62         echo ' </tr>';
63     } else if($i==4){
64         $arr2 = explode(",", $value);
65         echo ' <tr><td>Average</td>';
66         foreach ($arr2 as $val) {
67             echo ' <td>' . $val . ' </td>';
68         }
69         echo ' </tr>';
70     } else {
71         $arr2 = explode(",", $value);
72         echo ' <tr><td>Stardard_Deviation</td>';
73         foreach ($arr2 as $val) {
74             echo ' <td>' . $val . ' </td>';
75         }
76         echo ' </tr>
77
78         </tbody>
79         </table>
80         </div>
81         </div>
82         '
83         ;
84         $i=0;
85     }
86     $i+=1;
87 }
88
89 </div>
90 </div>
91 </div>
92 </div>
93 </body>
94 </html>
95 <script type="text/javascript">
96 $(document).ready(function(){
97
98     $( "#spark-data" ).submit(function( event ) {
99         window.open( 'localhost:50070/explorer.html#' + $( '#data_path' ).val(), '_blank' );
100     });
101 });
102 </script>

```

Listing B.4: nav.php

```

2 <div class="nav-wrapper_blue">
3 <ul id="nav-mobile" class="left_hide-on-med-and-down">
4 <li><a href="index.php" style="font-size:_40px;">Submit</a></li>
5 <li><a href="view.php" style="font-size:_40px;">Patern</a></li>
6 </ul>
7 </div>
8 </nav>

```

### Listing B.5: form.php

```

1 <form id="spark-form" class="col_m12" method="post" action="result.php" style="font-size:_20px;">
2 <div class="row">
3 <div class="input-field_col_m12_s12_black-text">
4 <input id="jar_path" type="text" name="jar_path" class="validate">
5 <label for="jar_path" >Spark JAR Path</label>
6 </div>
7 </div>
8 <div class="row">
9 <div class="input-field_col_m12_s12_black-text">
10 <input id="input_path" type="text" name="input_path" class="black-text">
11 <label for="input_path" >Input Path</label>
12 </div>
13 </div>
14 <div class="row">
15 <div class="input-field_col_m12_s12_black-text">
16 <input id="output_path" type="text" name="output_path" class="black-text">
17 <label for="output_path" >Output Path</label>
18 </div>
19 </div>
20
21 <div class="row">
22 <div class="input-field_col_m4_s12_black-text">
23 <input id="number_of_executor" type="number" name="executor_number" class="black-text" value="1" min="1" step="1" max=
    "100">
24 <label for="number_of_executor" >Number of Executor</label>
25 </div>
26 <div class="input-field_col_m4_s12_black-text">
27 <input id="executor_memory" type="number" name="executor_memory" class="black-text" value="1000" min="1000" step="100"
    >
28 <label for="executor_memory" >Executor Memory in mb</label>
29 </div>
30 <div class="input-field_col_m4_s12_black-text">
31 <input id="number_of_partition" type="number" name="number" class="black-text" value="1" min="1" step="1" max="200">
32 <label for="number_of_partition" >Number of Partition</label>
33 </div>
34 </div>
35 <div class="row">
36 <div class="input-field_col_m4_s12_black-text">
37 <input id="max_obj" type="number" name="max_obj" class="black-text" value="1" min="1" step="1" max="100">
38 <label for="max_obj" >Max Object</label>
39 </div>
40 <div class="input-field_col_m4_s12_black-text">
41 <select name="type">
42 <option value="0" >Single Linkage</option>
43 <option value="1" >Complete Linkage</option>
44 <option value="2" >Centroid Linkage</option>
45 </select>
46 </div>
47 <div class="input-field_col_m4_s12_black-text">
48 <input id="cut_off" type="number" name="cut_off" class="black-text" value="0.1" min="0.1" step="0.1" max="1">
49 <label for="cut_off" >Cut Off Distance</label>
50 </div>
51 </div>
52
53 <button class="btn_waves-effect_waves-light" type="submit" name="action">Submit
54 <i class="material-icons_right"></i>
55 </button>
56 </form>

```

### Listing B.6: list.php

```

1 <!DOCTYPE html>
2 <html>
3 <head>
4 <?php include 'head.php' ?>
5 <title>Spark Reduce Data App UI</title>
6 <style type="text/css">
7 </style>
8 </head>
9
10 <body>
11 <?php include 'nav.php' ?>
12 <div class="row">
13 <div class="col_m3_offset-m1">
14 <div class="row">
15 <div class="col_m12" style="margin-top:_20px;">
16 <div class="card_teal_lighten-5">
17 <div class="card-content_black-text">
18 <?php
19 $output = shell_exec('cd_/home/miebakso/hadoop-2.7.3&&_/bin/hadoop_fs_ls_'. $_POST['data_path']);
20 $sarr = explode("\n", $output);
21 $len = count($sarr)-2;
22 $i=0;
23 while($i<$len){
24 <echo "<a href='data.php?part=". $i."&path=". $_POST['data_path']."'>part-". $i."</a><br>";
25 $i=$i+1;

```

```

26         }
27     ?>
28 </div>
29 </div>
30 </div>
31 </div>
32 </div>
33 </div>
34 </body>
35 </html>
36 <script type="text/javascript">
37     $(document).ready(function(){
38
39         $("#spark-data").submit(function( event ) {
40             window.open( 'localhost:50070/explorer.html#' + $('#data_path').val(), '_blank' );
41         });
42     });
43 </script>

```

Listing B.7: result.php

```

1 <!DOCTYPE html>
2 <html>
3 <head>
4     <?php include 'head.php' ?>
5     <title>Spark Reduce Data App UI</title>
6     <style type="text/css">
7     </style>
8 </head>
9
10 <body>
11     <?php include 'nav.php' ?>
12     <div class="row">
13         <div class="col_m4_offset-m4">
14             <div class="row">
15                 <div class="col_m12" style="margin-top: 30px;">
16                     <div class="card_teal_lighten-5">
17                         <div class="card-content_black-text">
18                             <span class="card-title">Submit Successful</span>
19                             <div class="row">
20                                 <?php
21                                     $jar = $_POST['jar_path'];
22                                     $input = $_POST['input_path'];
23                                     $output = $_POST['output_path'];
24                                     $executor_number = $_POST['executor_number'];
25                                     $executor_memory = $_POST['executor_memory'];
26                                     $partition = $_POST['number'];
27                                     $max_obj = $_POST['max_obj'];
28                                     $type = $_POST['type'];
29                                     $cutoff = $_POST['cut_off'];
30
31                                     $output = shell_exec('cd $_SPARK_HOME_&&_./bin/spark-submit_-class_main.scala.Main_-master_yarn_/home/
32                                         miebakso/IdeaProjects/BigData/target/scala-2.11/bigdata_2.11-0.1.jar');
33                                     ?>
34                                 </div>
35                             </div>
36                         </div>
37                     </div>
38                 </div>
39             </div>
40 </body>
41 </html>

```

Listing B.8: view.php

```

1 <!DOCTYPE html>
2 <html>
3 <head>
4     <?php include 'head.php' ?>
5     <title>Spark Reduce Data App UI</title>
6     <style type="text/css">
7     </style>
8 </head>
9
10 <body>
11     <?php include 'nav.php' ?>
12     <div class="row">
13         <div class="col_m6_offset-m3">
14             <div class="row">
15                 <div class="col_m12" style="margin-top: 200px;">
16                     <div class="card_teal_lighten-5">
17                         <div class="card-content_black-text">
18                             <span class="card-title">Explore HDFS Directory</span>
19                             <div class="row">
20                                 <form id="spark-data" class="col_m12" action="list.php" method="post">
21                                     <div class="row">
22                                         <div class="input-field_col_m12_black-text">
23                                             <input id="data_path" type="text" name="data_path" class="validate">
24                                             <label for="data_path">HDFS data path</label>
25                                         </div>
26                                     </div>
27
28                                     <button class="btn_waves-effect_waves-light" type="submit" name="action">Submit
29                                     <i class="material-icons_right"></i>
30                                 </button>

```



```
31 |         </form>
32 |     </div>
33 | </div>
34 | </div>
35 | </div>
36 | </div>
37 | </body>
38 | </html>
39 | <script type="text/javascript">
40 |     $(document).ready(function(){
41 |
42 |         $( "#spark-data" ).submit(function( event ) {
43 |             window.open('localhost:50070/explorer.html#'+$('#data_path').val(), '_blank');
44 |         });
45 |     });
46 | </script>
```