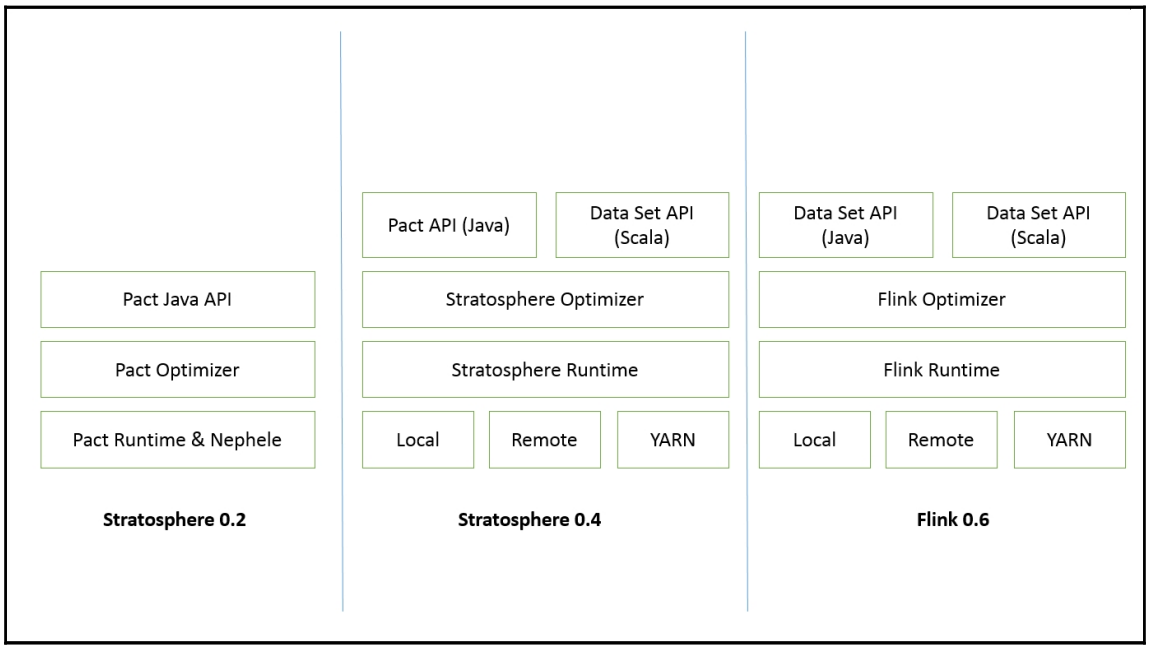
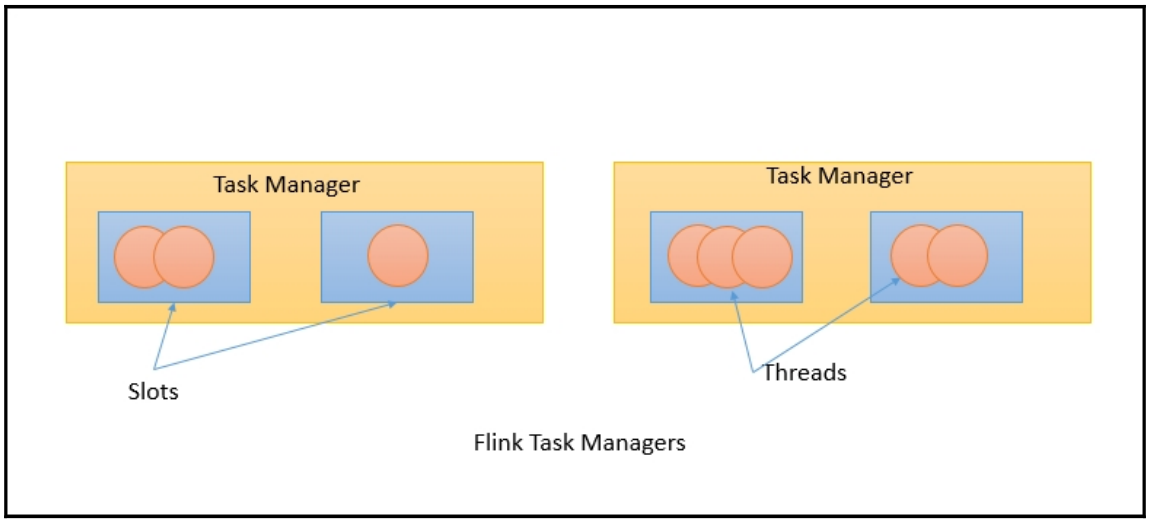
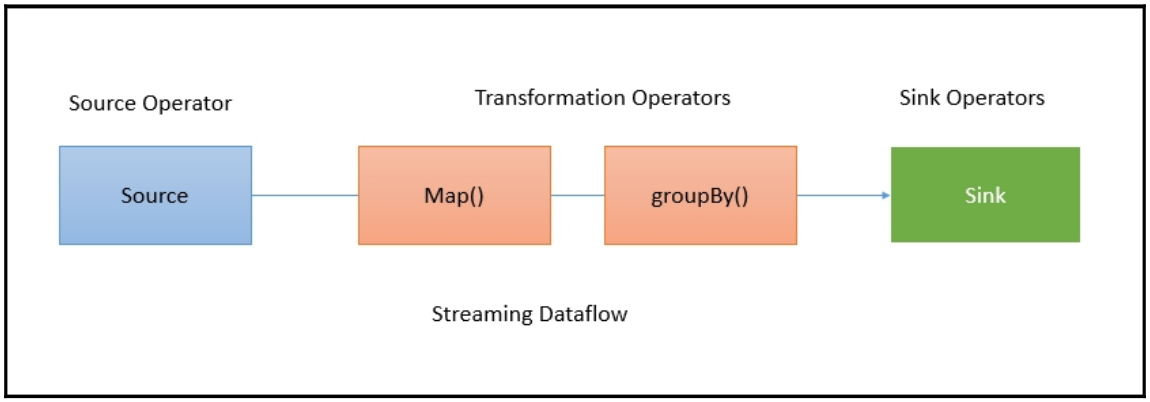
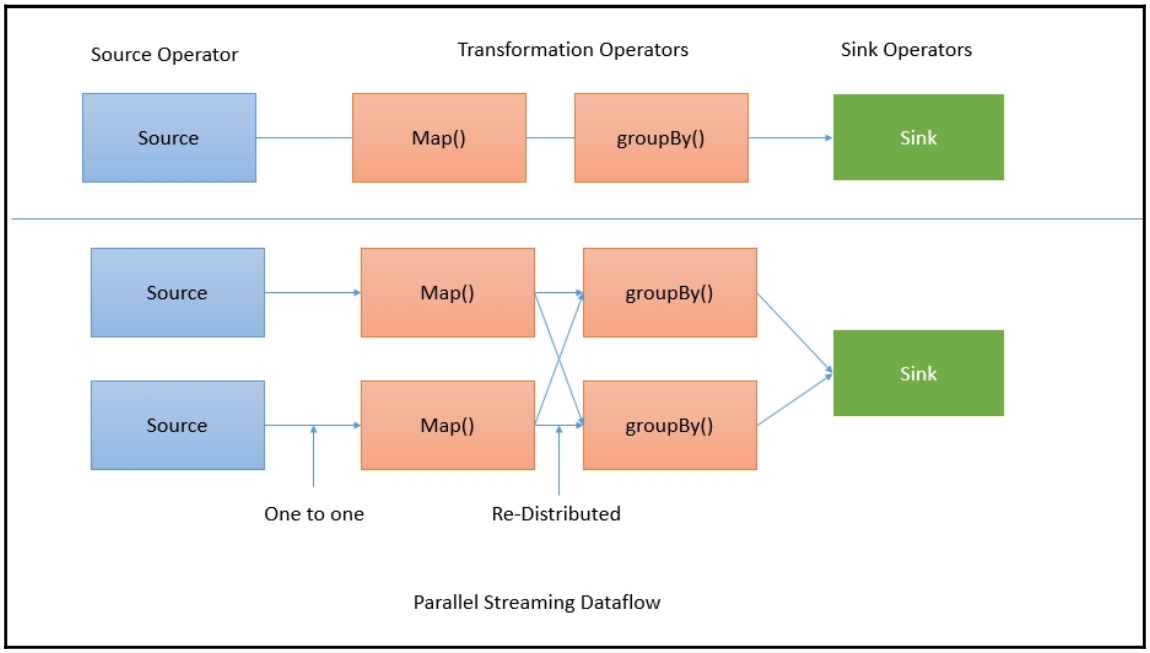
Chapter 1: Introduction to Apache Flink





```
val text =  env.readTextFile("input.txt"))       // Source

val counts = text.flatMap { _.toLowerCase.split("\\W+") filter { _.nonEmpty } }
      .map { (_, 1) }
      .groupBy(0)
      .sum(1)                                    // Transformation

counts.writeAsCsv("output.txt", "\n", " ")       // Sink
```

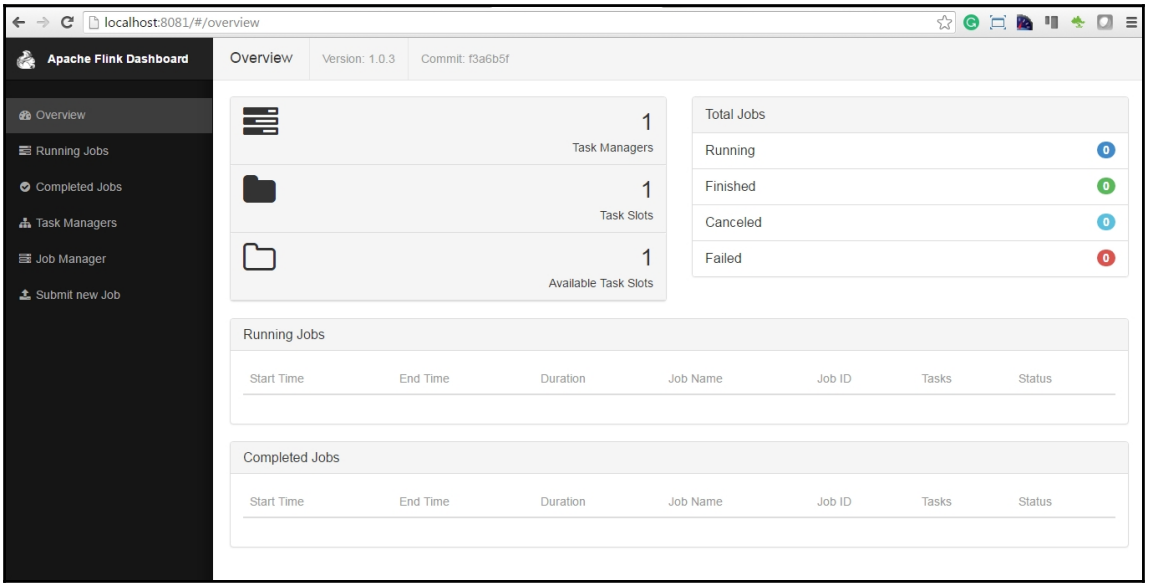


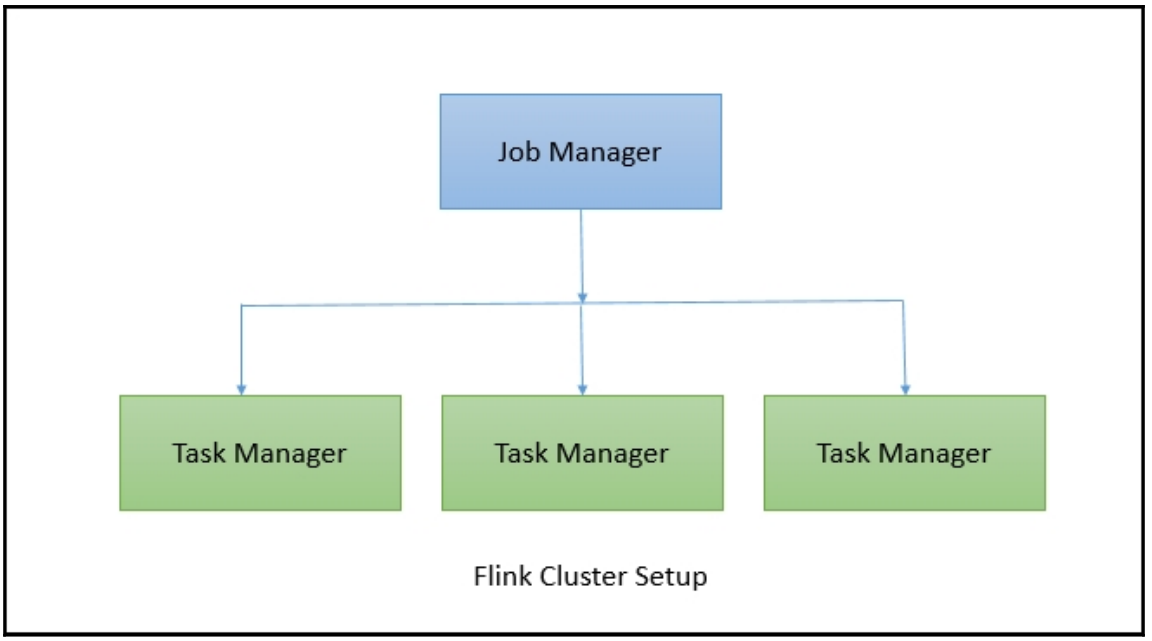


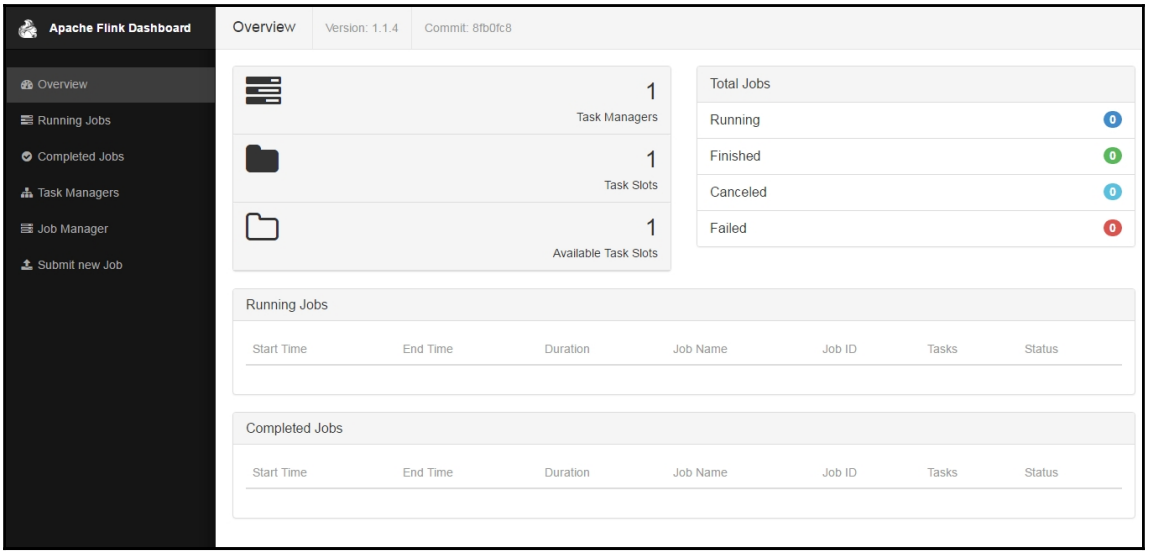
Binaries

	Scala 2.10	Scala 2.11
Hadoop 1.2.1	Download	
Hadoop 2.3.0	Download	Download
Hadoop 2.4.1	Download	Download
Hadoop 2.6.0	Download	Download
Hadoop 2.7.0	Download	Download

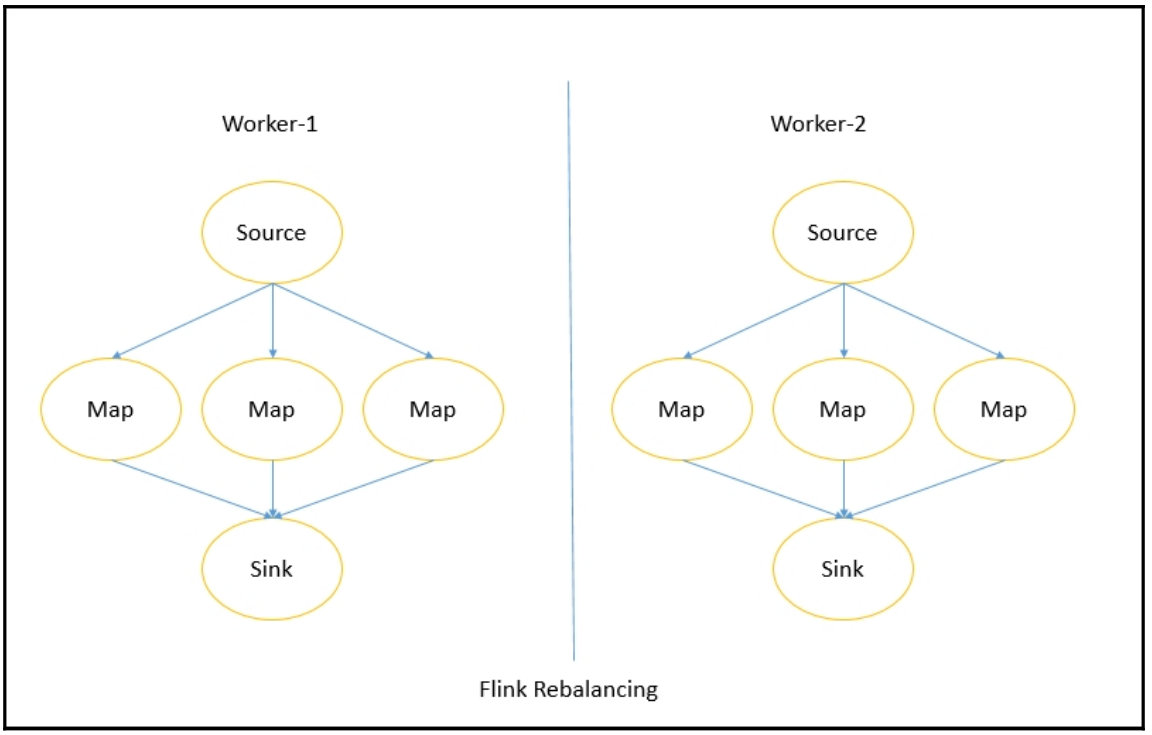
```
D:\>java -version
java version "1.8.0_92"
Java(TM) SE Runtime Environment (build 1.8.0_92-b14)
Java HotSpot(TM) 64-Bit Server VM (build 25.92-b14, mixed mode)
```

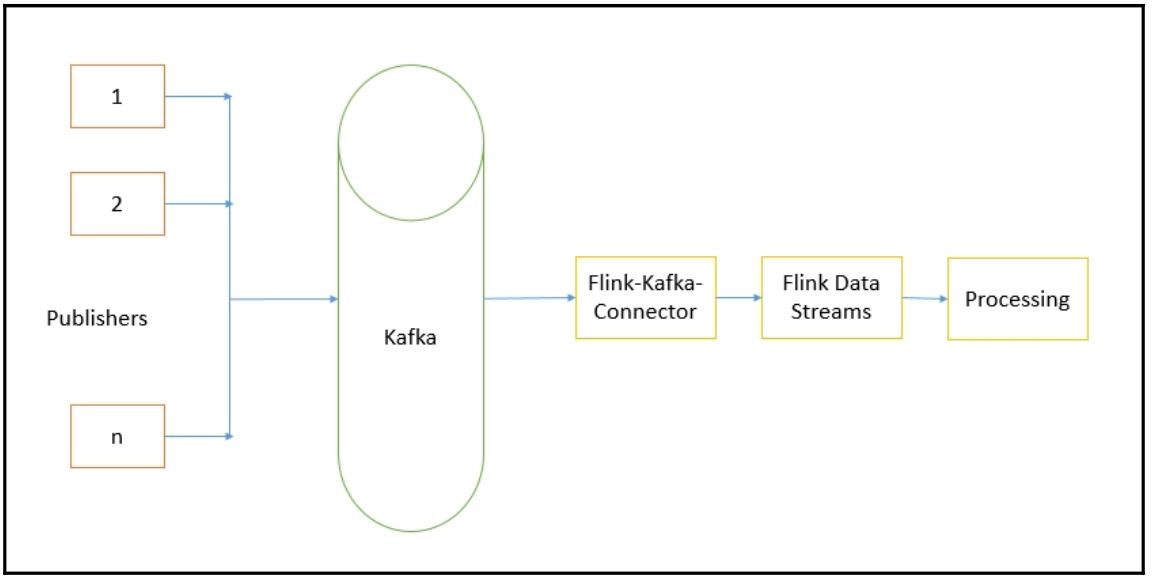






Chapter 2: Data Processing Using the DataStream API


Worker-1
Source
Map
Map
Map
Sink
Worker-2
Source
Map
Map
Map
Sink
Flink Rebalancing


1
2
Publishers
n
Kafka
Flink-Kafka-Connector
Flink Data Streams
Processing

HadoopTrendingTopics

Test OAuth

Details | Settings | Keys and Access Tokens | Permissions

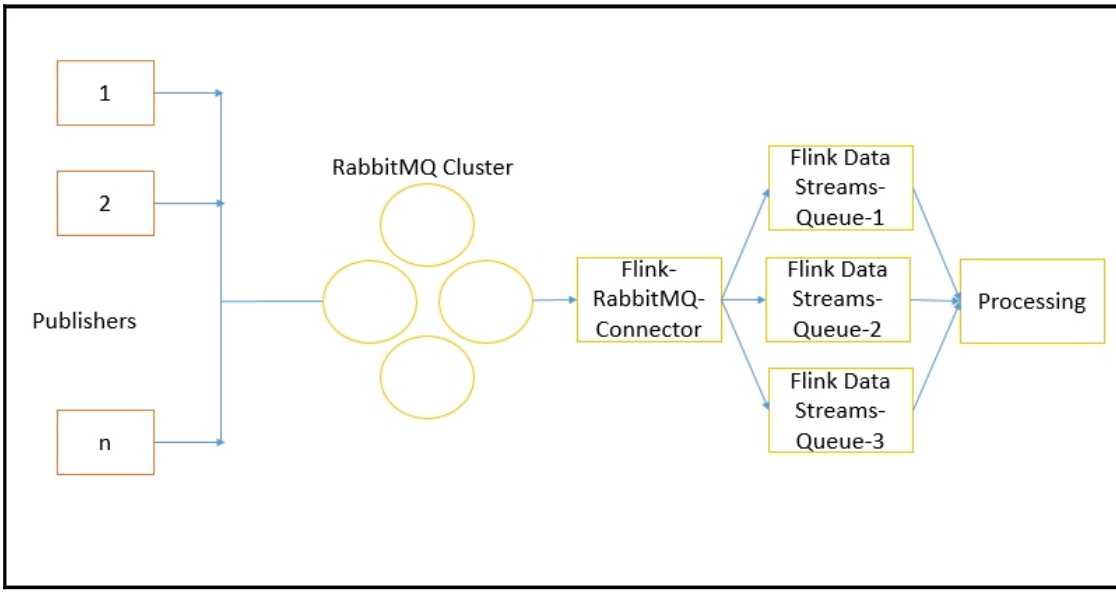
Application Settings

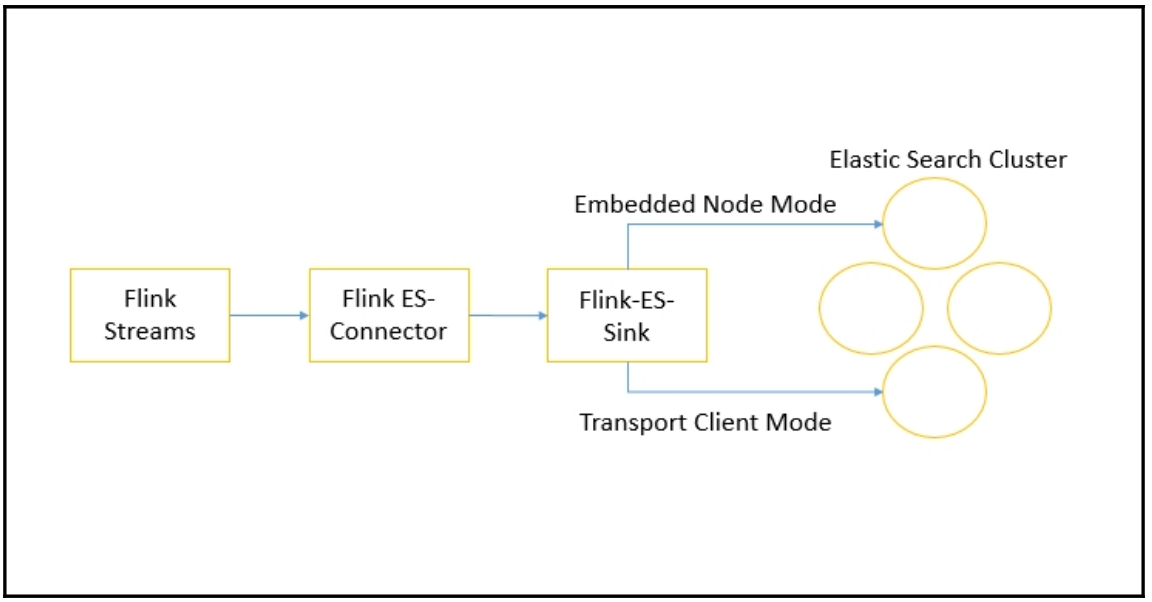
Keep the "Consumer Secret" a secret. This key should never be human-readable in your application.

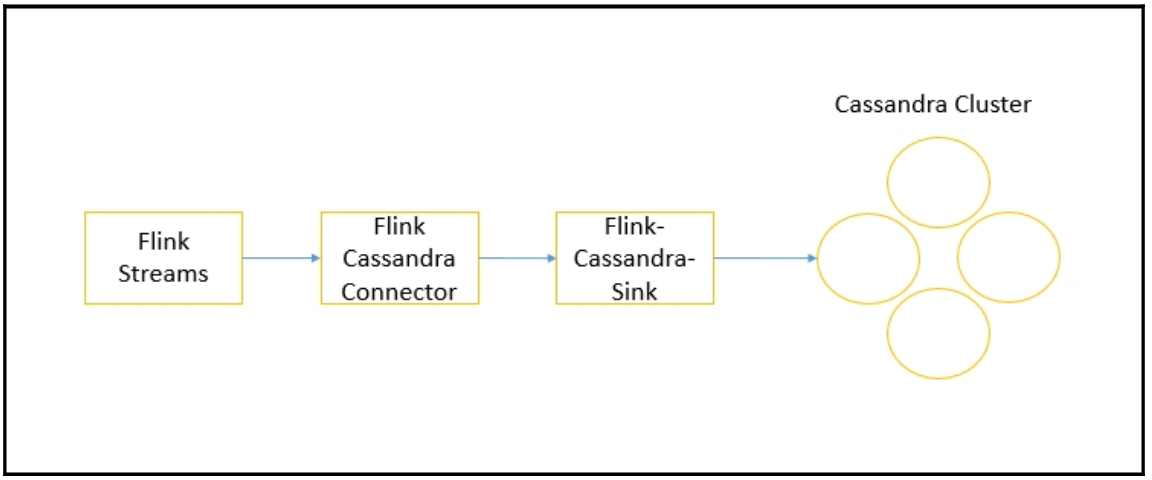
Consumer Key (API Key)	tP686cWPcJAdCsuF4Alv
Consumer Secret (API Secret)	HkP6yxcByJqHAypaGGII75Npcu5GkbrGMgOQRqIT8
Access Level	Read-only (modify app permissions)
Owner	HadoopTutorials
Owner ID	2825680861

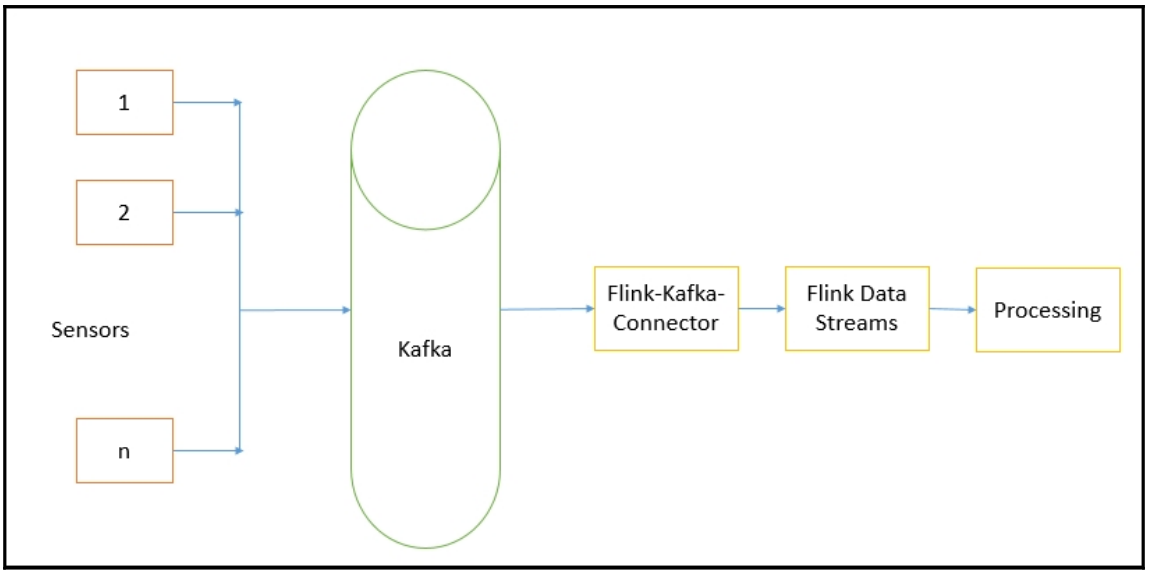
Application Actions

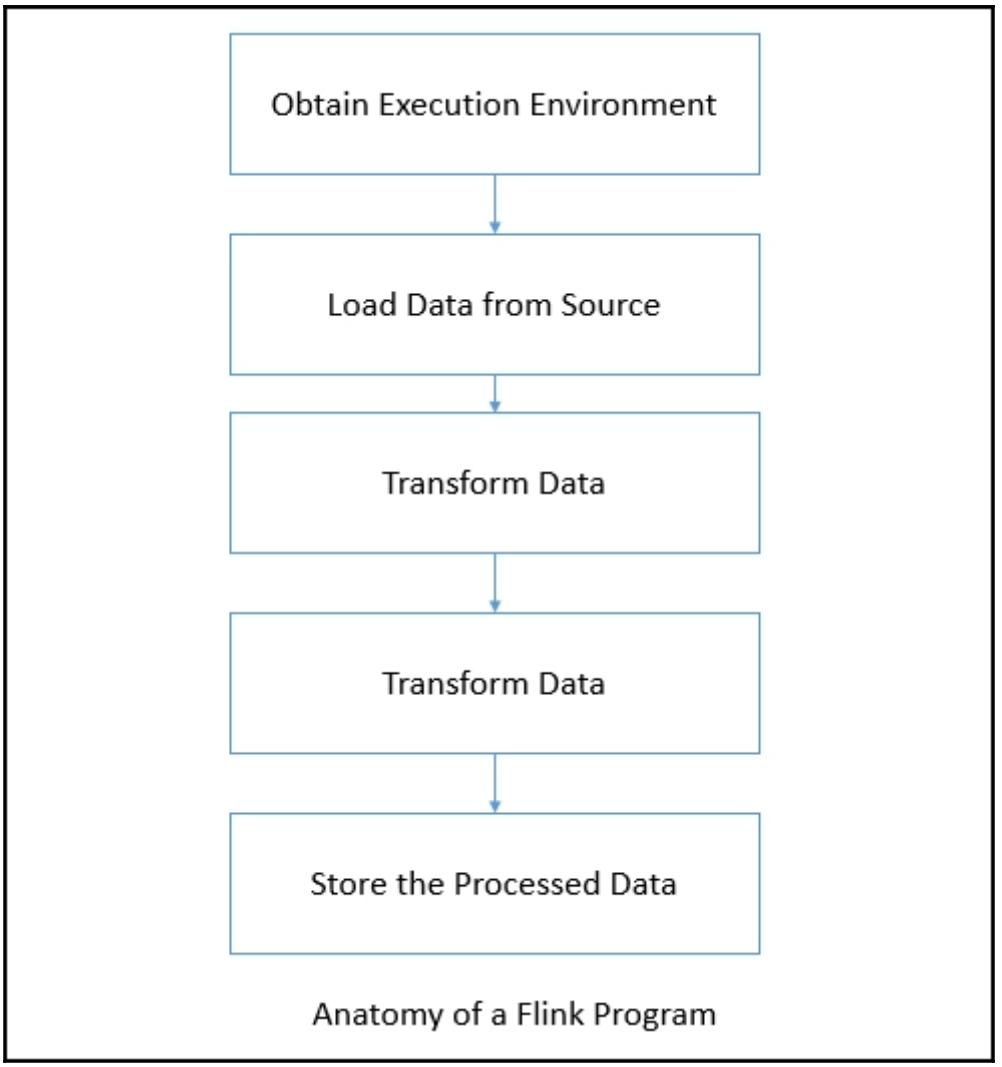
Regenerate Consumer Key and Secret | Change App Permissions




Elastic Search Cluster
Embedded Node Mode
Flink Streams
Flink ES-Connector
Flink-ES-Sink
Transport Client Mode

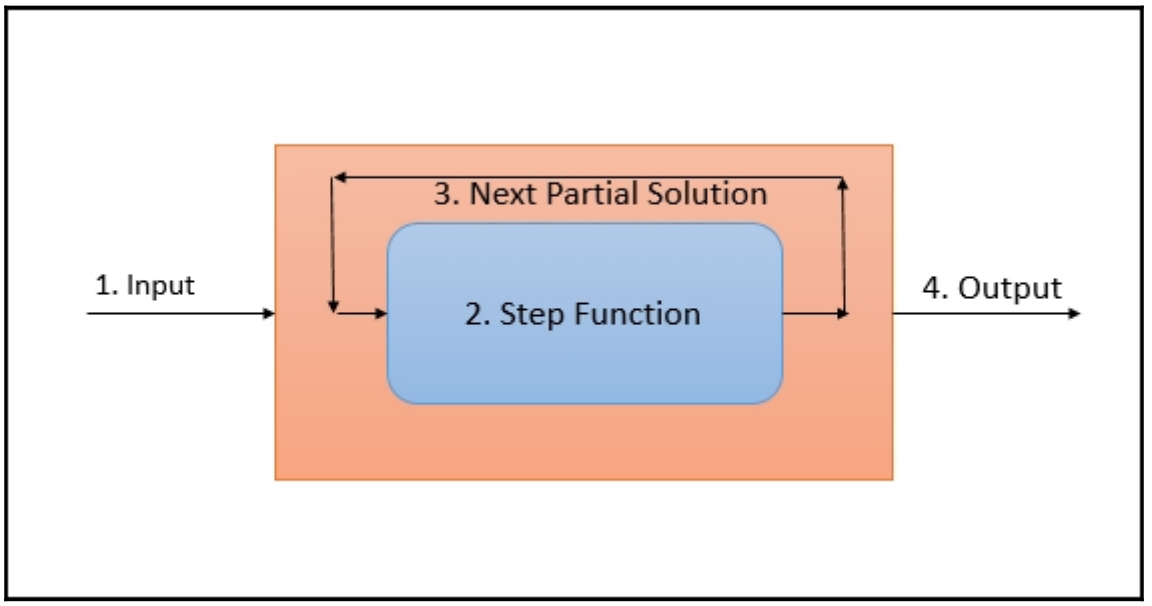

Cassandra Cluster
Flink Streams
Flink Cassandra Connector
Flink-Cassandra-Sink

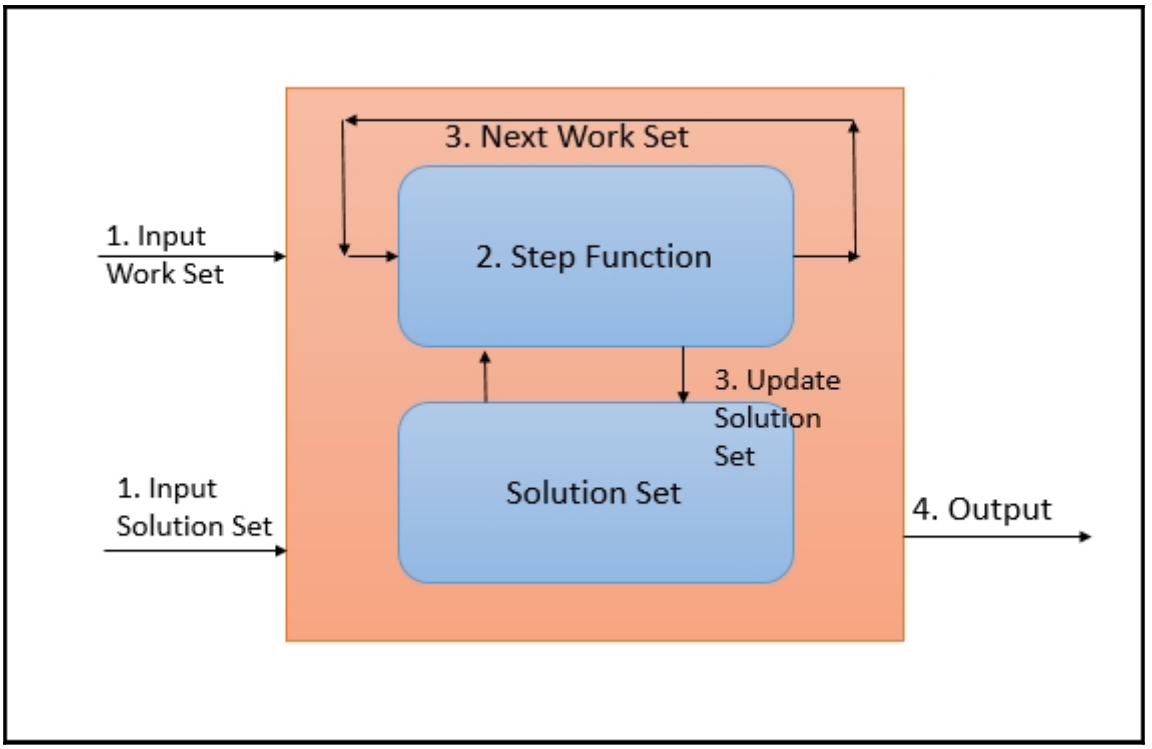

1
2
Sensors
n
Kafka
Flink-Kafka-Connector
Flink Data Streams
Processing



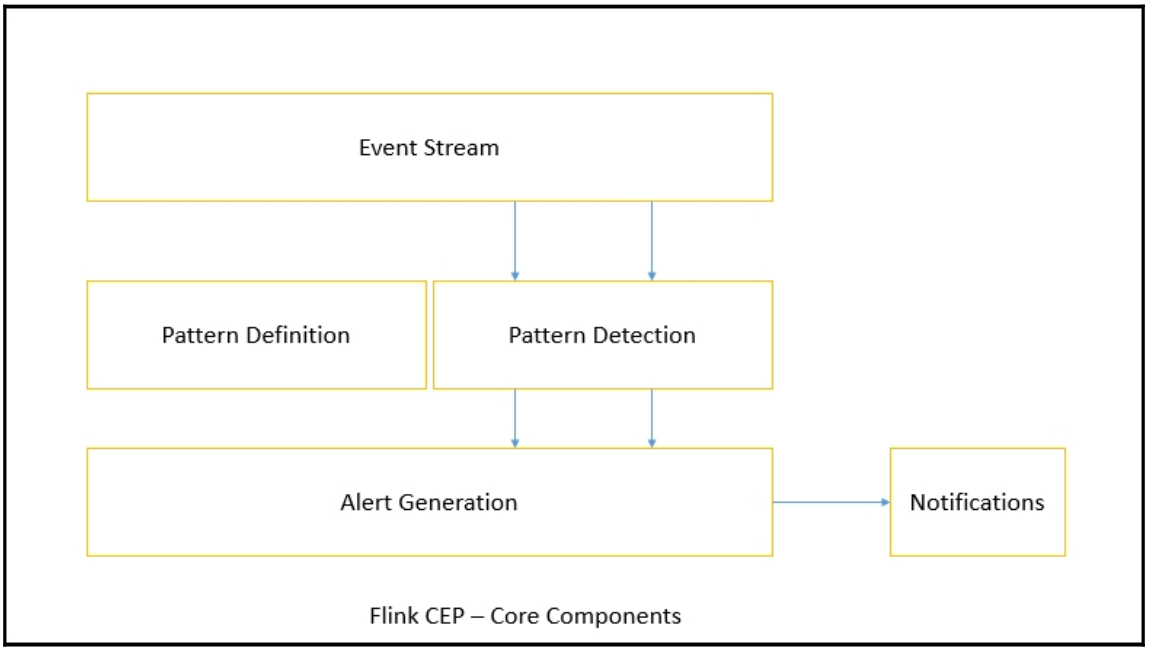
Anatomy of a Flink Program

Chapter 3: Data Processing Using the Batch Processing API

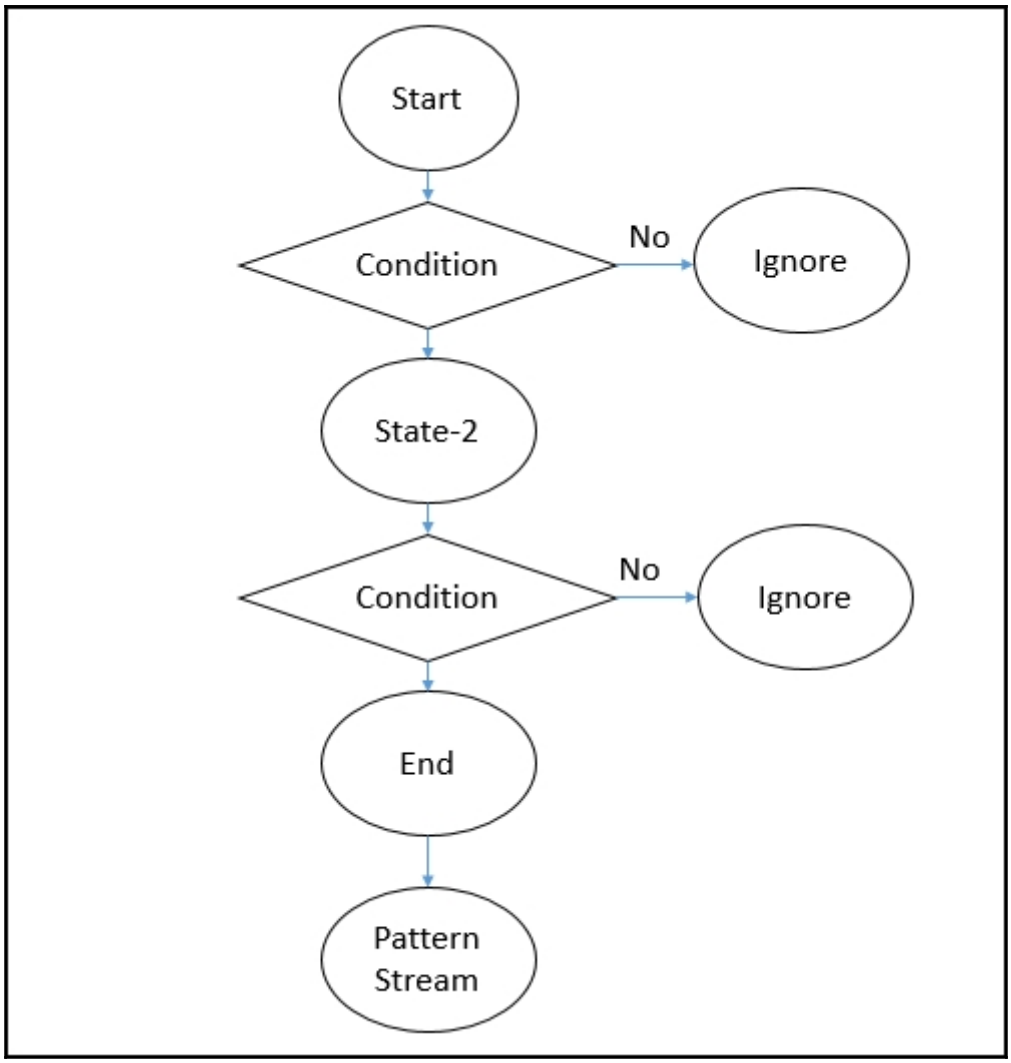


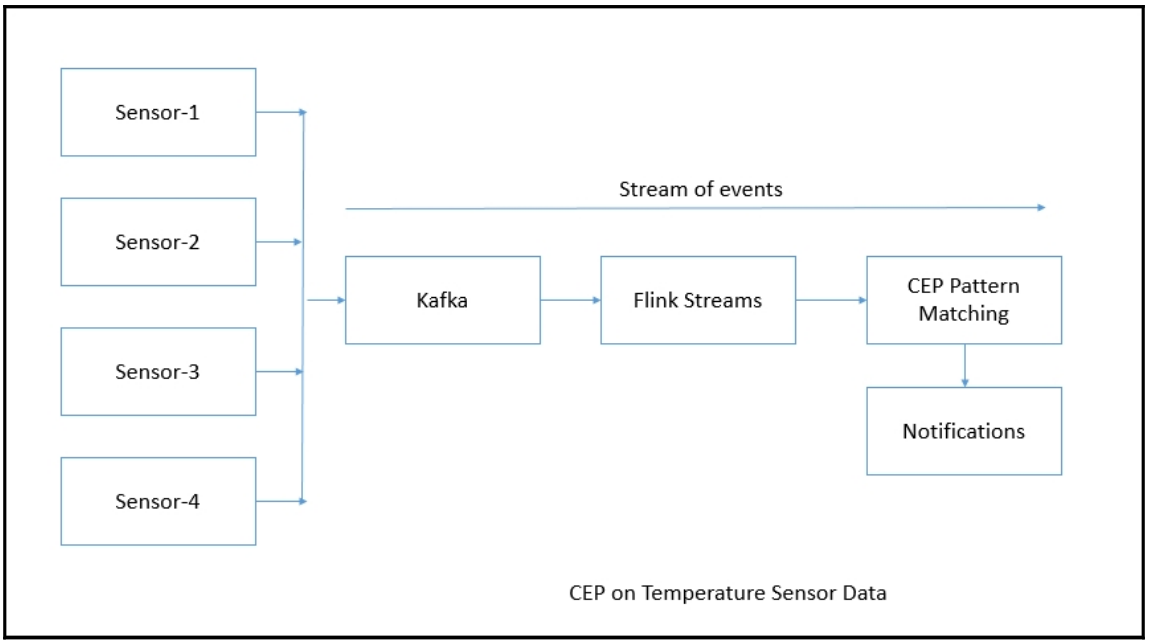

3. Next Work Set
1. Input
Work Set
2. Step Function
3. Update
Solution
Set
1. Input
Solution Set
Solution Set
4. Output

Chapter 5: Complex Event Processing

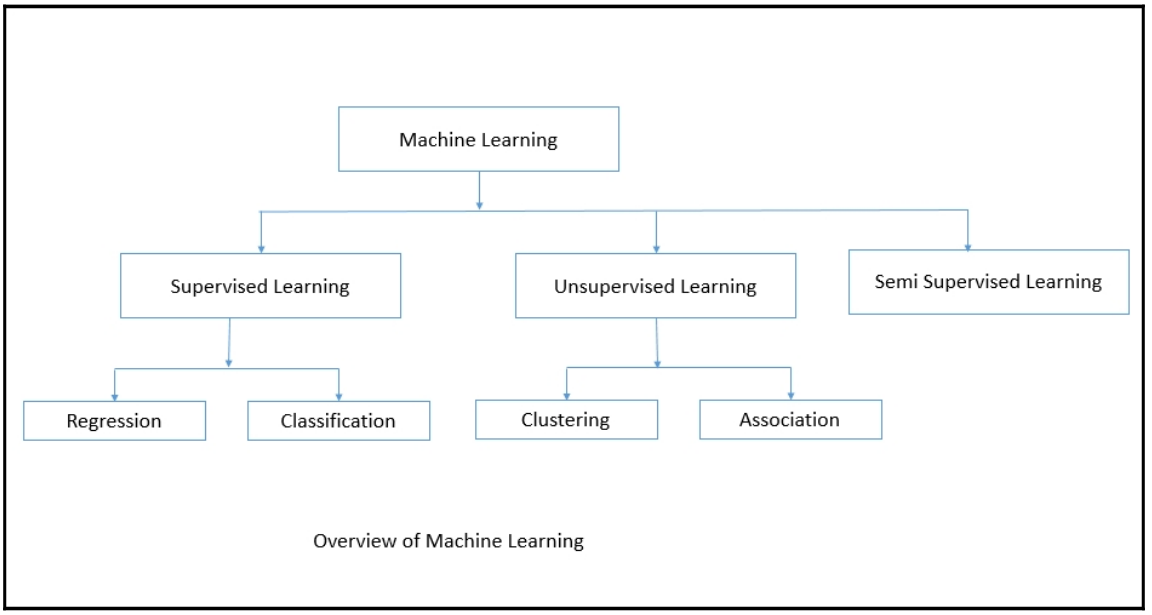


Flink CEP – Core Components


Start
Condition
No
Ignore
State-2
Condition
No
Ignore
End
Pattern
Stream

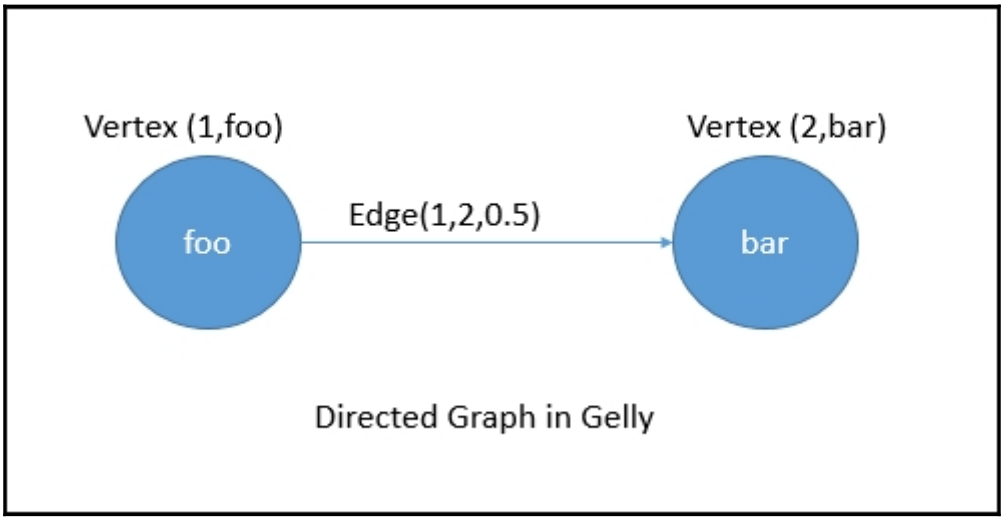

Sensor-1
Sensor-2
Sensor-3
Sensor-4
Stream of events
Kafka
Flink Streams
CEP Pattern Matching
Notifications
CEP on Temperature Sensor Data

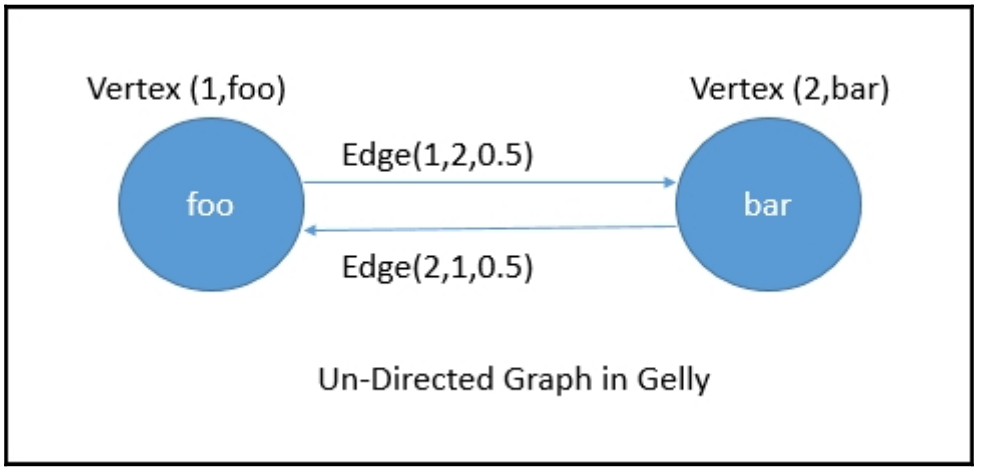
Chapter 6: Machine Learning Using FlinkML

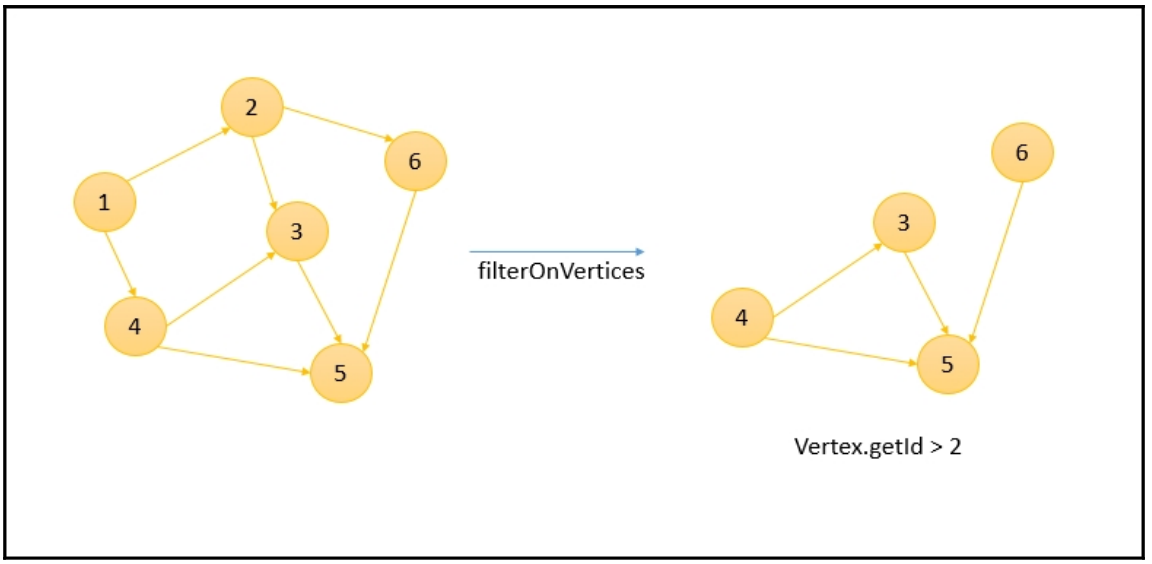


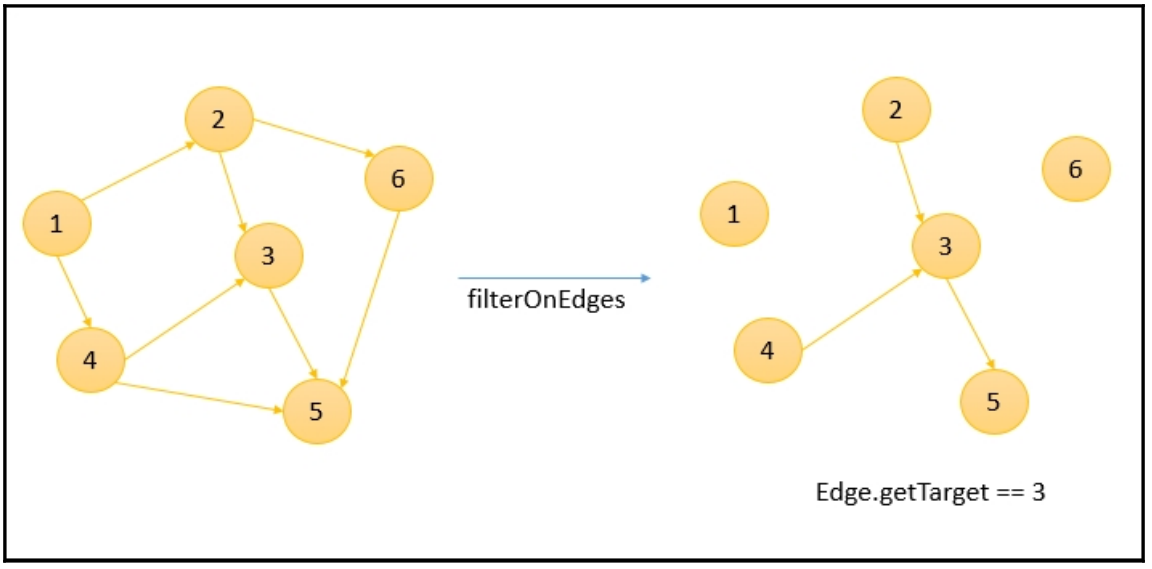
Overview of Machine Learning

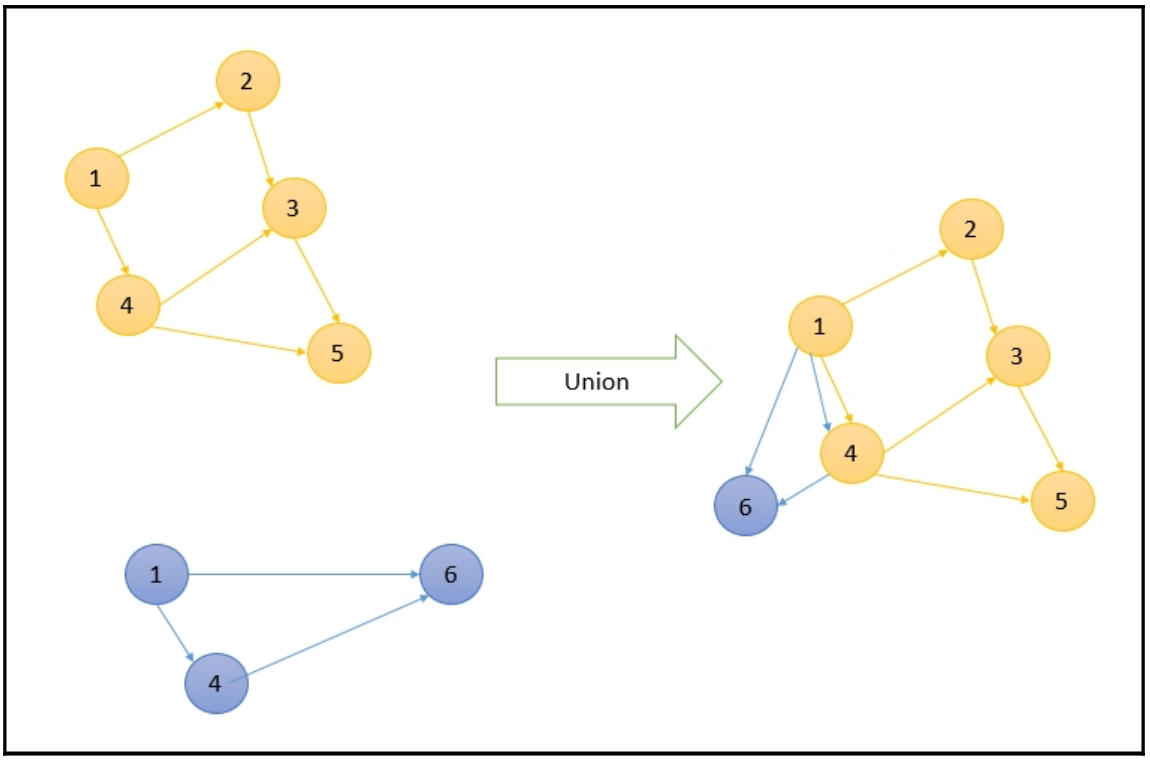
Chapter 7: Flink Graph API – Gelly

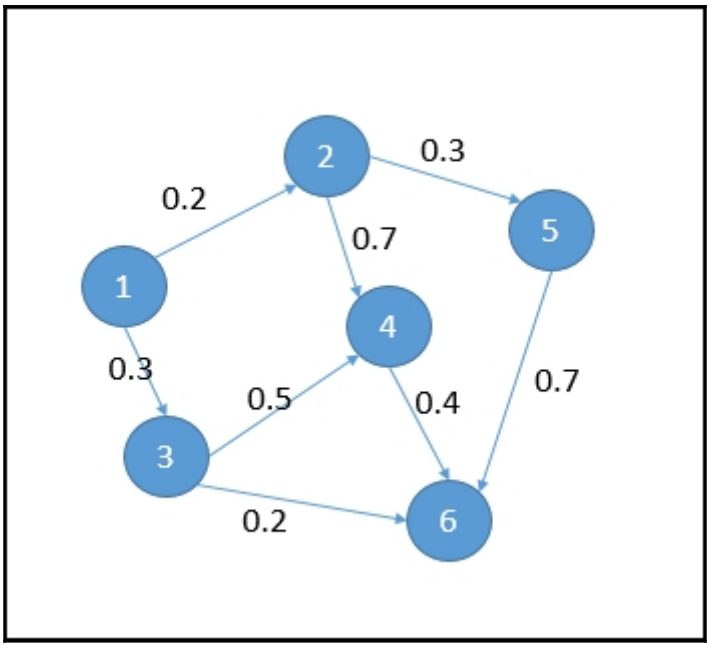


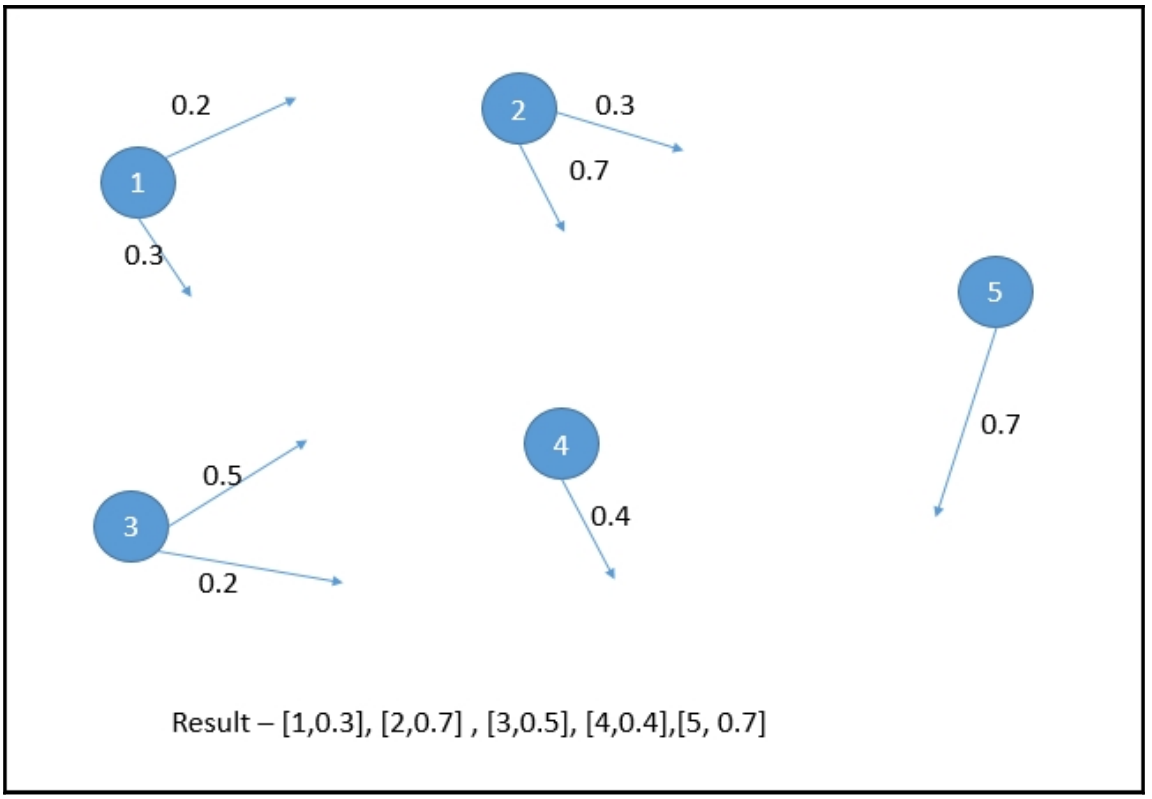



2
6
1
3
4
5
filterOnVertices
6
3
4
5
Vertex.getId > 2

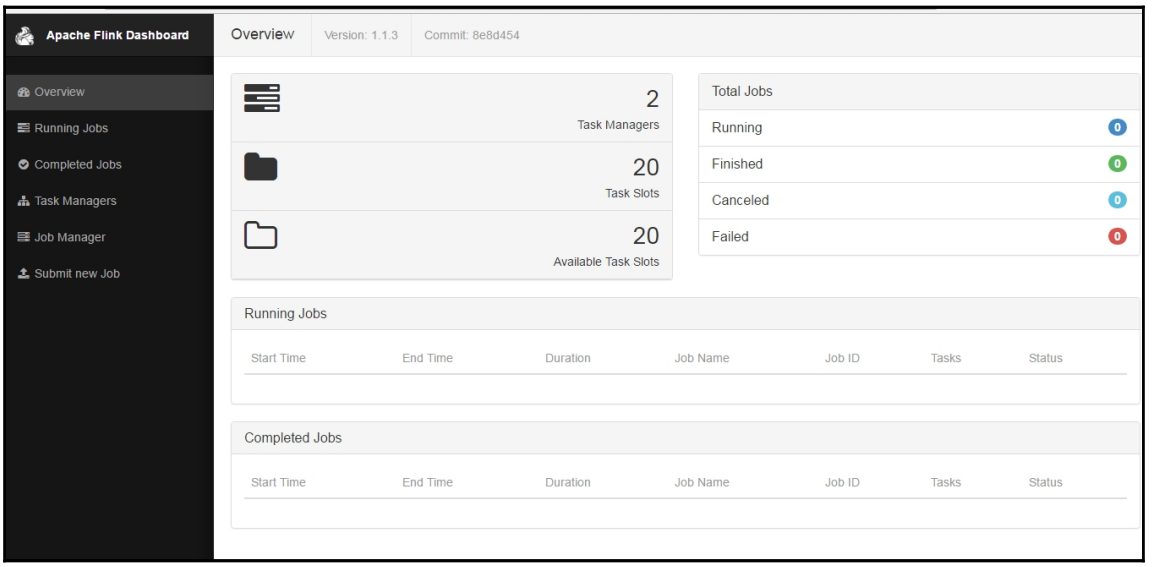

2
6
1
3
4
5
filterOnEdges
2
6
1
3
4
5
Edge.getTarget == 3

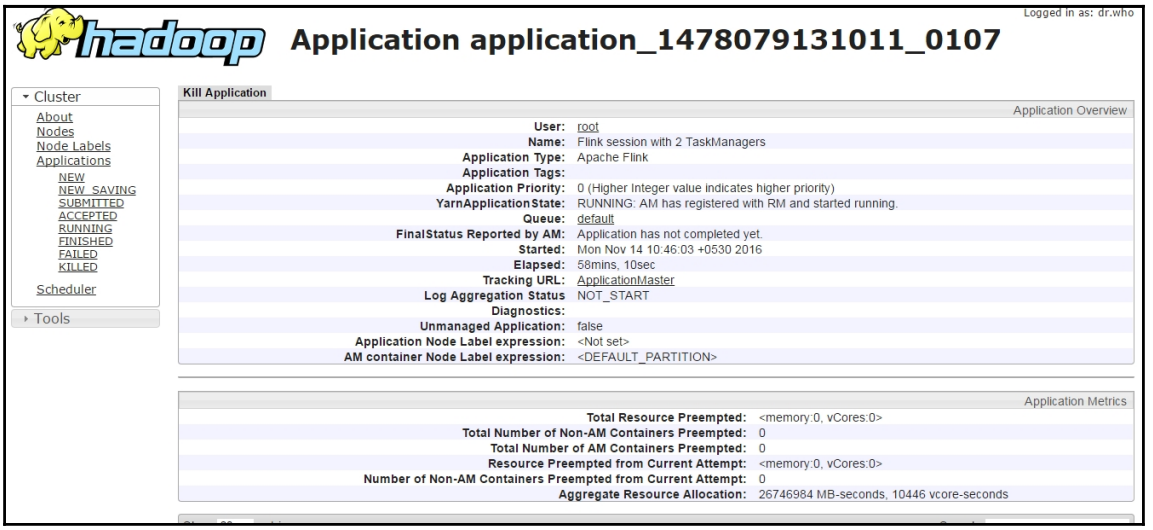

2
1
3
4
5
Union
2
1
3
4
6
5
1
6
4

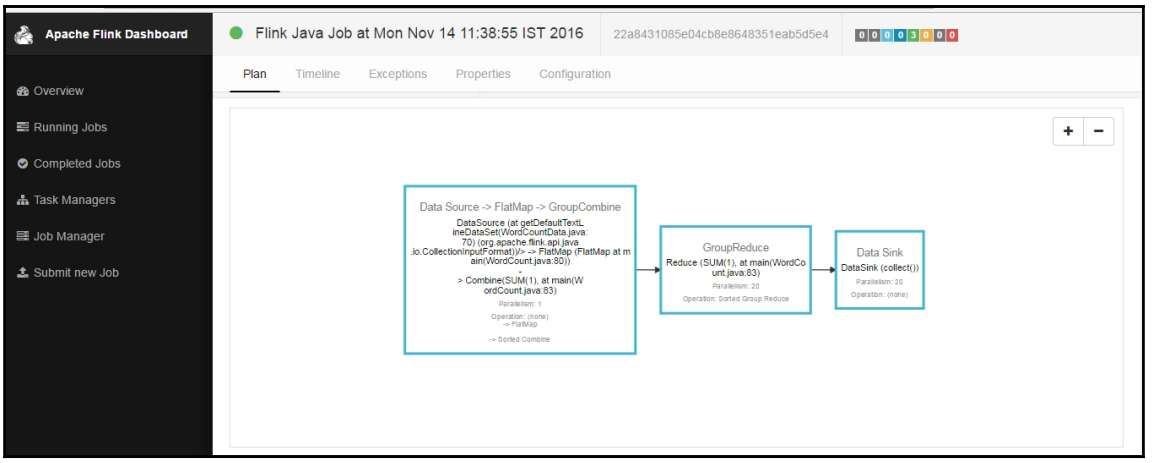

2
0.3
0.2
5
0.7
1
4
0.3
0.5
0.4
0.7
3
6
0.2


0.2
1
0.3
2
0.3
0.7
5
0.7
4
0.4
3
0.5
0.2
Result – [1,0.3], [2,0.7] , [3,0.5], [4,0.4],[5, 0.7]

Chapter 8: Distributed Data Processing with Flink and Hadoop

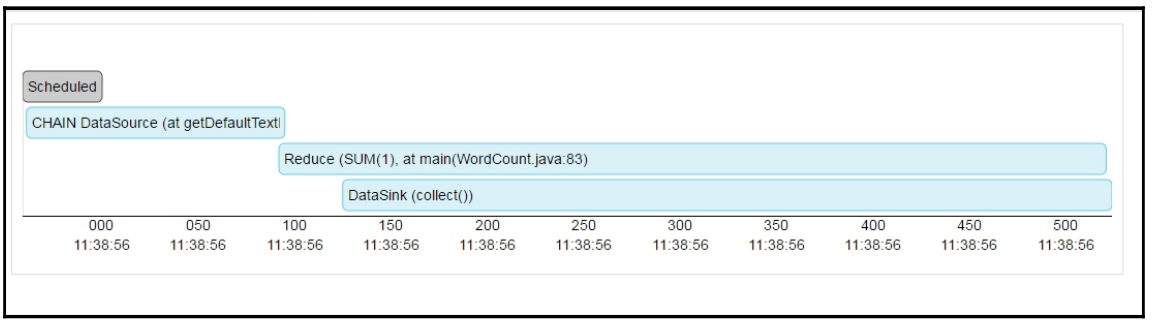






Subtasks | TaskManagers | Accumulators | Checkpoints

Start Time	End Time	Duration	Name	Bytes received	Records received	Bytes sent	Records sent	Tasks	Status
2016-11-14, 11:38:55	2016-11-14, 11:38:56	134ms	CHAIN DataSource (at getDefaultTextLineDataSet(WordCountData.java:70) (org.apache.flink.api.java.io.CollectionInputFormat)) -> FlatMap (FlatMap at main(WordCount.java:80)) -> Combine(SUM(1), at main(WordCount.java:83)	0 B	0	1.66 KB	170	0 0 0 1 0 0 0	FINISHED
2016-11-14, 11:38:56	2016-11-14, 11:38:56	429ms	Reduce (SUM(1), at main(WordCount.java:83)	1.66 KB	170	1.66 KB	170	0 0 0 20 0 0 0	FINISHED
2016-11-14, 11:38:56	2016-11-14, 11:38:56	399ms	DataSink (collect())	1.66 KB	170	0 B	0	0 0 0 20 0 0 0	FINISHED



Application Overview

User:	root
Name:	Flink Application: org.apache.flink.examples.java.wordcount.WordCount
Application Type:	Apache Flink
Application Tags:	
Application Priority:	0 (Higher Integer value indicates higher priority)
YarnApplicationState:	FINISHED
Queue:	default
FinalStatus Reported by AM:	SUCCEEDED
Started:	Mon Nov 14 12:00:16 +0530 2016
Elapsed:	9sec
Tracking URL:	History
Log Aggregation Status	SUCCEEDED
Diagnostics:	Flink YARN Client requested shutdown
Unmanaged Application:	false
Application Node Label expression:	<Not set>
AM container Node Label expression:	<DEFAULT_PARTITION>

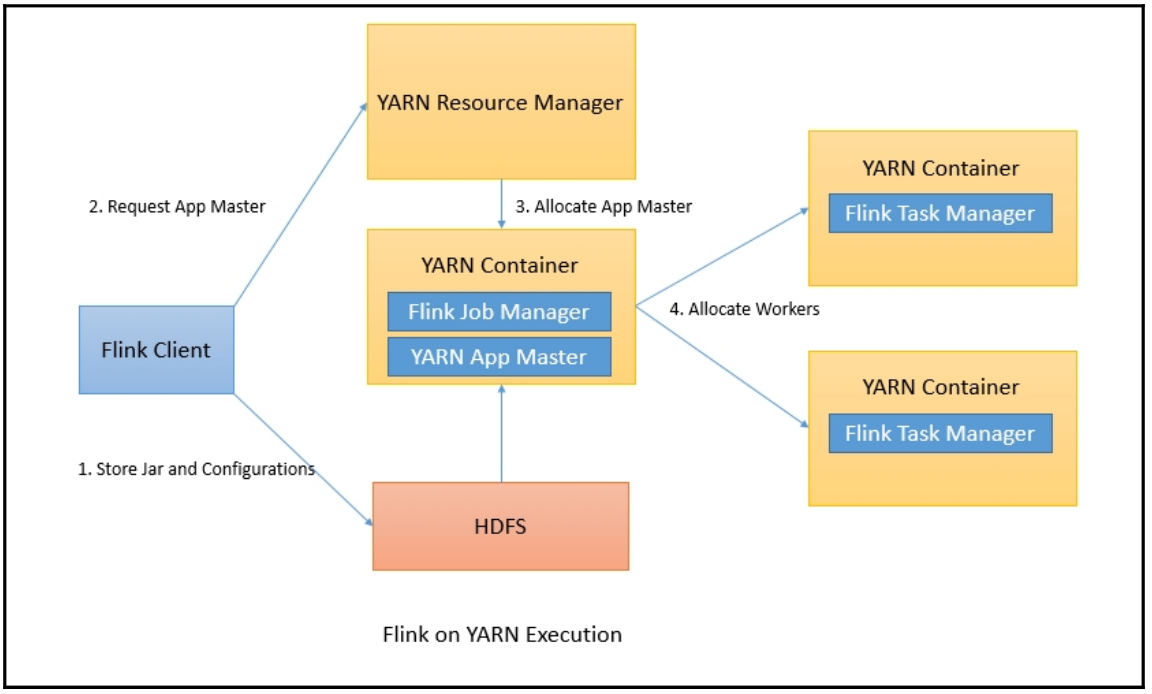
Application Metrics

Total Resource Preempted:	<memory:0, vCores:0>
Total Number of Non-AM Containers Preempted:	0
Total Number of AM Containers Preempted:	0
Resource Preempted from Current Attempt:	<memory:0, vCores:0>
Number of Non-AM Containers Preempted from Current Attempt:	0
Aggregate Resource Allocation:	48130 MB-seconds, 17 vcore-seconds

Show 20 entries Search:

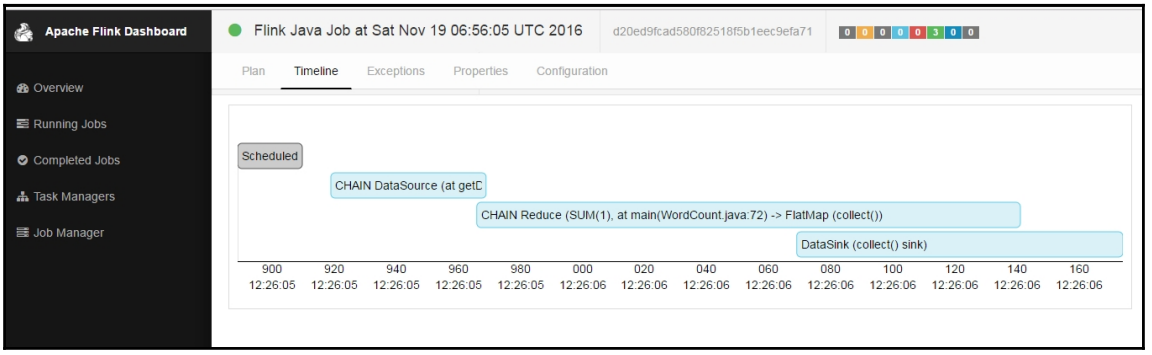
Attempt ID	Started	Node	Logs	Blacklisted Nodes
appattempt_1478079131011_0108_000001	Mon Nov 14 12:00:16 +0550 2016	http://hdpdev005.pune-in0145.slb.com:8042	Logs	N/A

Showing 1 to 1 of 1 entries First Previous 1 Next Last

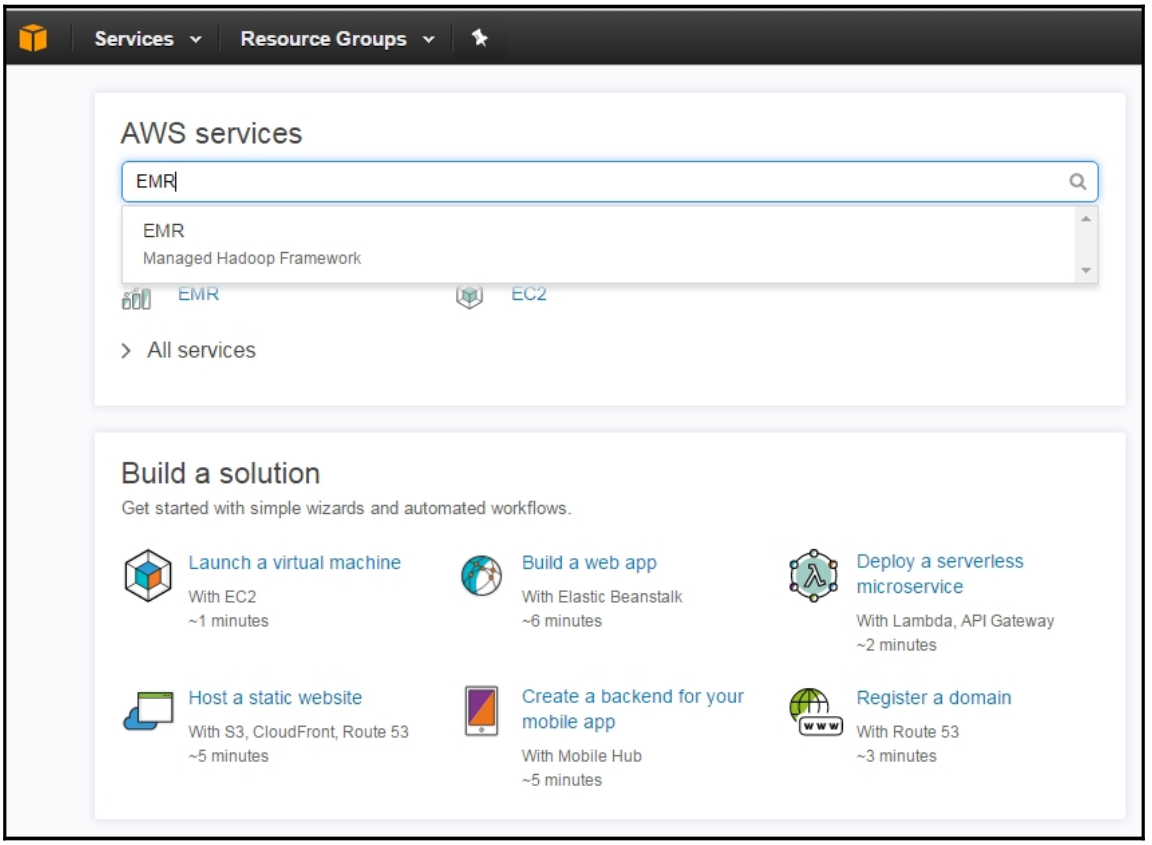


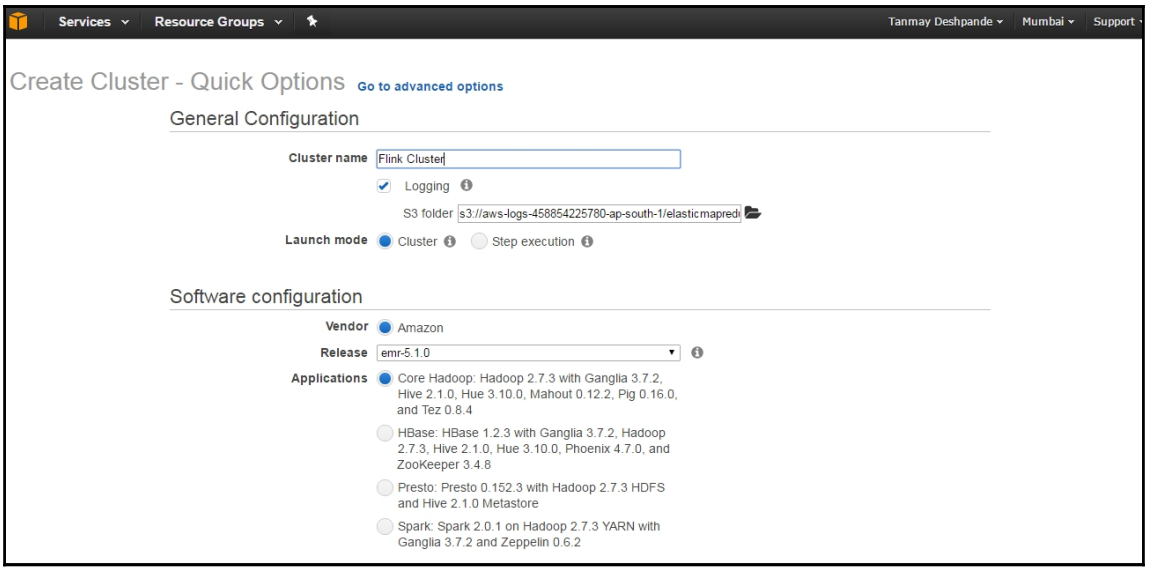
Flink on YARN Execution

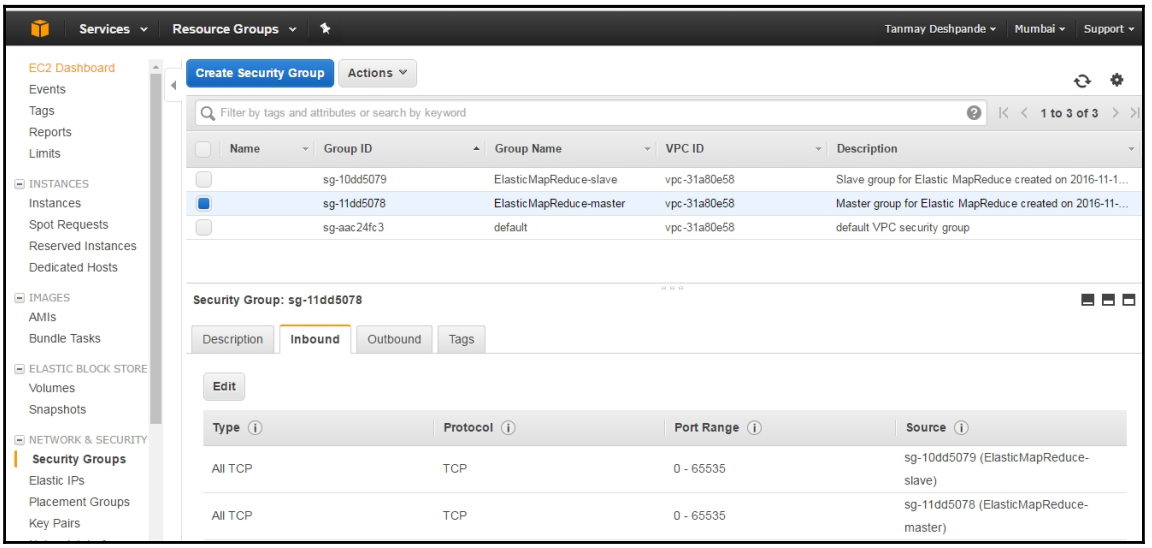
Chapter 9: Deploying Flink on Cloud



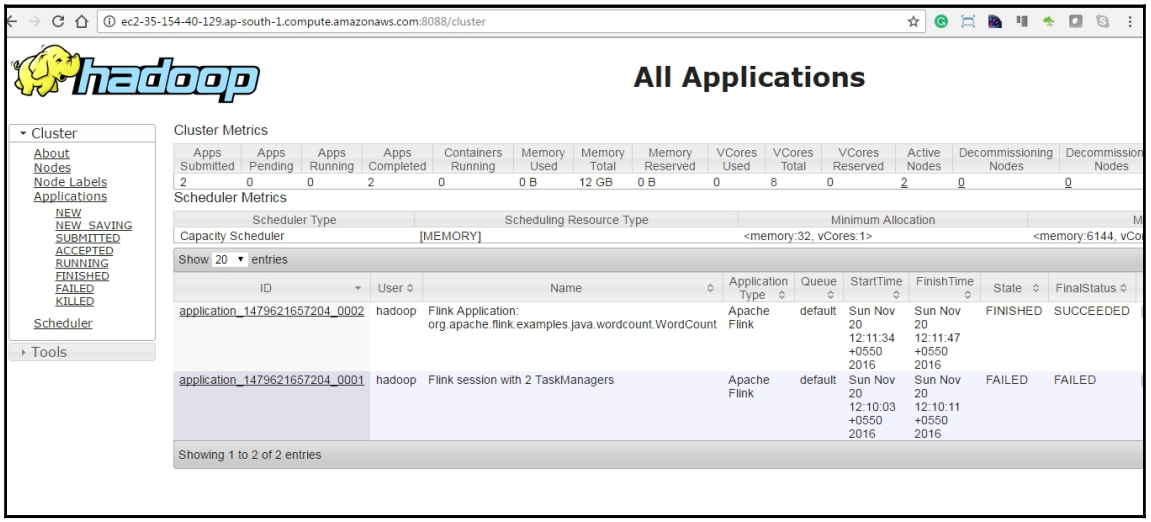
```
Sat Nov 19 08:34:05 UTC 2016: Using local tmp dir for staging files: /tmp/bdutil-20161119-083405-GdK
Sat Nov 19 08:34:05 UTC 2016: Using custom environment-variable file(s): bdutil_env.sh extensions/flink/flink_env.sh
Sat Nov 19 08:34:05 UTC 2016: Reading environment-variable file: ./bdutil_env.sh
Sat Nov 19 08:34:05 UTC 2016: Reading environment-variable file: extensions/flink/flink_env.sh
Sat Nov 19 08:34:05 UTC 2016: No explicit GCE_MASTER_MACHINE_TYPE provided; defaulting to value of GCE_MACHINE_TYPE: n1-standard-2
Delete cluster with following settings?
      CONFIGBUCKET='bdutil-flink-bucket'
      PROJECT='
      GCE_IMAGE='https://www.googleapis.com/compute/v1/projects/debian-cloud/global/images/backports-debian-7-wheezy-v20160531'
      GCE_ZONE='europe-west1-c'
      GCE_NETWORK='default'
      GCE_TAGS='bdutil'
      PREEMPTIBLE_FRACTION=0.0
      PREFIX='hadoop'
      NUM_WORKERS=2
      MASTER_HOSTNAME='hadoop-m'
      WORKERS='hadoop-w-0 hadoop-w-1'
      BDUTIL_GCS_STAGING_DIR='gs://bdutil-flink-bucket/bdutil-staging/hadoop-m'
      (y/n)
```

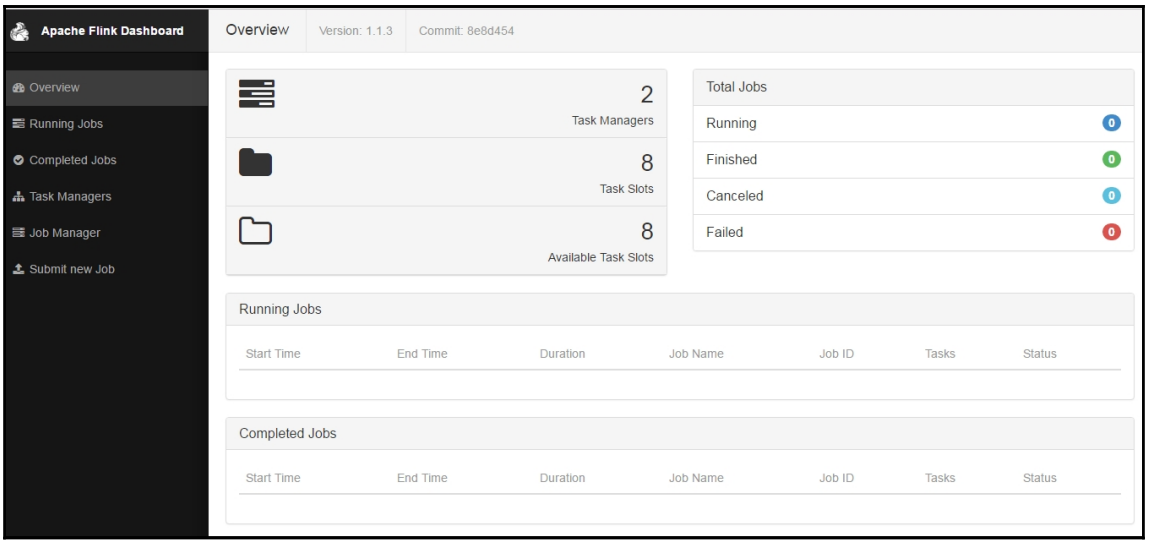


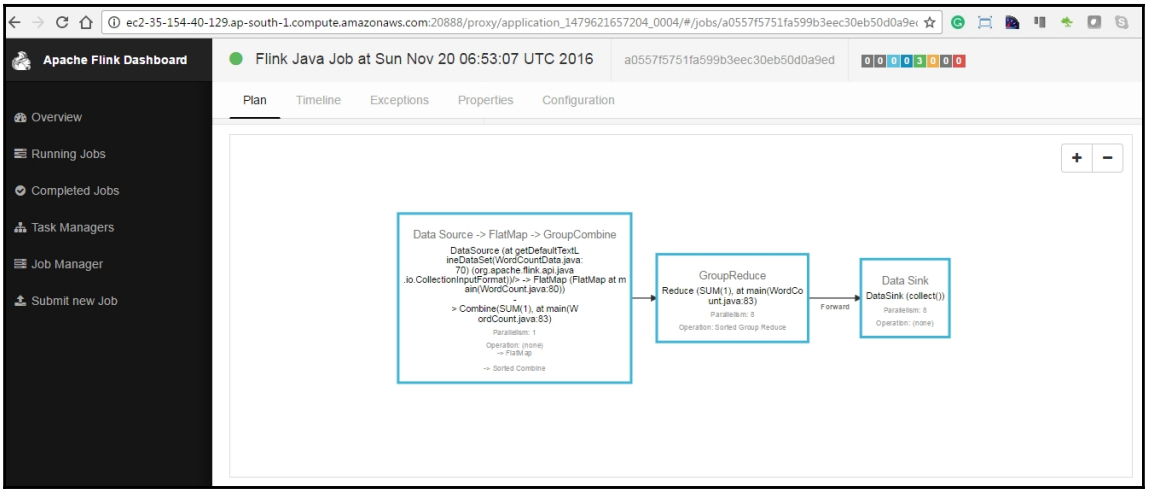

Services
Resource Groups
Tanmay Deshpande
Mumbai
Support
Create Cluster - Quick Options
Go to advanced options
General Configuration
Cluster name
Flink Cluster
Logging
S3 folder
s3://aws-logs-458854225780-ap-south-1/elasticmapred
Launch mode
Cluster
Step execution
Software configuration
Vendor
Amazon
Release
emr-5.1.0
Applications
Core Hadoop: Hadoop 2.7.3 with Ganglia 3.7.2, Hive 2.1.0, Hue 3.10.0, Mahout 0.12.2, Pig 0.16.0, and Tez 0.8.4
HBase: HBase 1.2.3 with Ganglia 3.7.2, Hadoop 2.7.3, Hive 2.1.0, Hue 3.10.0, Phoenix 4.7.0, and ZooKeeper 3.4.8
Presto: Presto 0.152.3 with Hadoop 2.7.3 HDFS and Hive 2.1.0 Metastore
Spark: Spark 2.0.1 on Hadoop 2.7.3 YARN with Ganglia 3.7.2 and Zeppelin 0.6.2

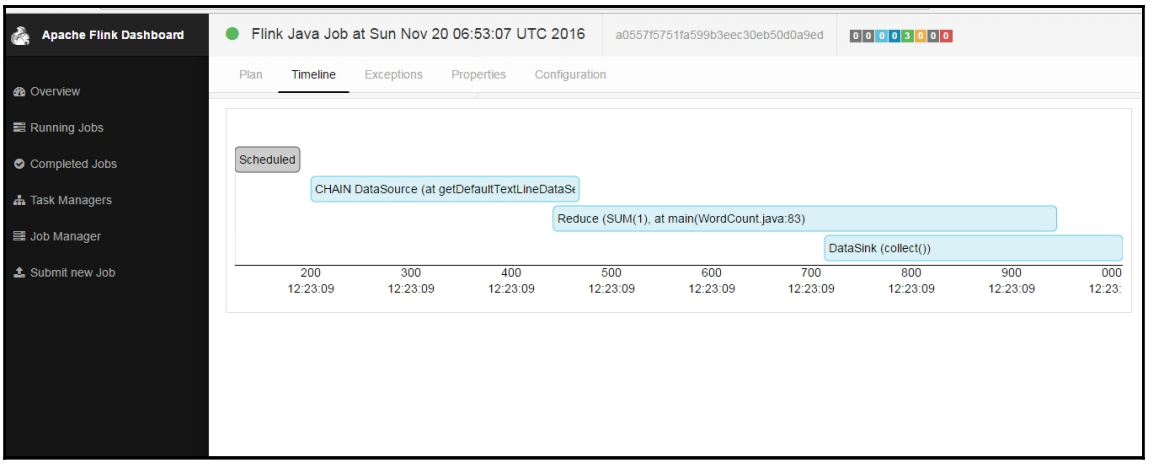

Services
Resource Groups
Tanmay Deshpande
Mumbai
Support
EC2 Dashboard
Events
Tags
Reports
Limits
INSTANCES
Instances
Spot Requests
Reserved Instances
Dedicated Hosts
IMAGES
AMIs
Bundle Tasks
ELASTIC BLOCK STORE
Volumes
Snapshots
NETWORK & SECURITY
Security Groups
Elastic IPs
Placement Groups
Key Pairs
Create Security Group
Actions
Filter by tags and attributes or search by keyword
1 to 3 of 3
Name
Group ID
Group Name
VPC ID
Description
sg-10dd5079
ElasticMapReduce-slave
vpc-31a80e58
Slave group for Elastic MapReduce created on 2016-11-1...
sg-11dd5078
ElasticMapReduce-master
vpc-31a80e58
Master group for Elastic MapReduce created on 2016-11-...
sg-aac24fc3
default
vpc-31a80e58
default VPC security group
Security Group: sg-11dd5078
Description
Inbound
Outbound
Tags
Edit
Type
Protocol
Port Range
Source
All TCP
TCP
0 - 65535
sg-10dd5079 (ElasticMapReduce-slave)
All TCP
TCP
0 - 65535
sg-11dd5078 (ElasticMapReduce-master)

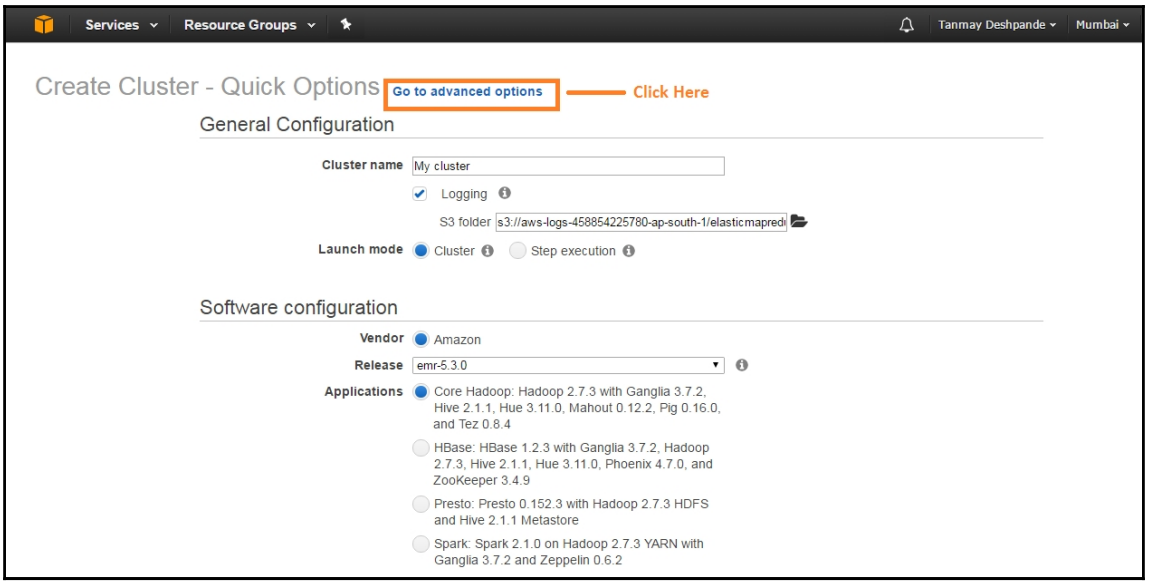
```
Using username "hadoop".
Authenticating with public key "imported-openssh-key"
Last login: Sun Nov 20 06:24:42 2016

       __|  __|_  )
       _|  (     /   Amazon Linux AMI
      ___|\___|___|

https://aws.amazon.com/amazon-linux-ami/2016.09-release-notes/
8 package(s) needed for security, out of 13 available
Run "sudo yum update" to apply all updates.

EEEEEEEEEEEEEEEEEEE MMMMMMMM           MMMMMMMM RRRRRRRRRRRRRRR
E::::::::::::::::::E M:::::::M         M:::::::M R::::::::::::::R
EE:::::EEEEEEEEE:::E M::::::::M       M::::::::M R:::::RRRRRR:::::R
  E::::E       EEEEE M:::::::::M     M:::::::::M RR::::R      R::::R
  E::::E             M::::::M:::M   M:::M::::::M   R:::R      R::::R
  E:::::EEEEEEEEEE   M:::::M M:::M M:::M M:::::M   R:::RRRRRR:::::R
  E::::::::::::::E   M:::::M  M:::M:::M  M:::::M   R:::::::::::RR
  E:::::EEEEEEEEEE   M:::::M   M:::::M   M:::::M   R:::RRRRRR:::::R
  E::::E             M:::::M    M:::M    M:::::M   R:::R      R::::R
  E::::E       EEEEE M:::::M     MMM     M:::::M   R:::R      R::::R
EE:::::EEEEEEEE::::E M:::::M             M:::::M   R:::R      R::::R
E::::::::::::::::::E M:::::M             M:::::M RR::::R      R::::R
EEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-2-68 ~]$
```

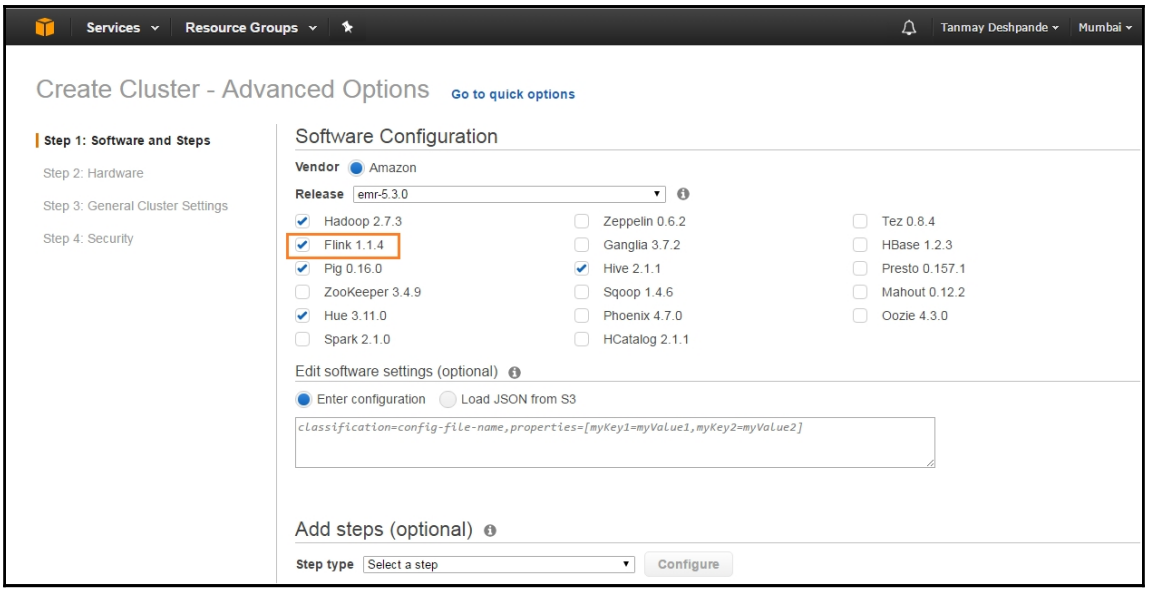



Apache Flink Dashboard
Overview
Version: 1.1.3
Commit: 8e8d454
Overview
Running Jobs
Completed Jobs
Task Managers
Job Manager
Submit new Job
2
Task Managers
8
Task Slots
8
Available Task Slots
Total Jobs
Running 0
Finished 0
Canceled 0
Failed 0
Running Jobs
Start Time
End Time
Duration
Job Name
Job ID
Tasks
Status
Completed Jobs
Start Time
End Time
Duration
Job Name
Job ID
Tasks
Status


ec2-35-154-40-129.ap-south-1.compute.amazonaws.com:20888/proxy/application_1479621657204_0004/#/jobs/a0557f5751fa599b3eec30eb50d0a9ed
Apache Flink Dashboard
Flink Java Job at Sun Nov 20 06:53:07 UTC 2016
a0557f5751fa599b3eec30eb50d0a9ed
Plan
Timeline
Exceptions
Properties
Configuration
Overview
Running Jobs
Completed Jobs
Task Managers
Job Manager
Submit new Job
Data Source -> FlatMap -> GroupCombine
GroupReduce
Reduce (SUM(1), at main(WordCount.java:83)
Forward
Data Sink
DataSink (collect())


Apache Flink Dashboard
Flink Java Job at Sun Nov 20 06:53:07 UTC 2016
a0557f5751fa599b3eec30eb50d0a9ed
Plan
Timeline
Exceptions
Properties
Configuration
Overview
Running Jobs
Completed Jobs
Task Managers
Job Manager
Submit new Job
Scheduled
CHAIN DataSource (at getDefaultTextLineDataSe
Reduce (SUM(1), at main(WordCount.java:83)
DataSink (collect())
200
12:23:09
300
12:23:09
400
12:23:09
500
12:23:09
600
12:23:09
700
12:23:09
800
12:23:09
900
12:23:09

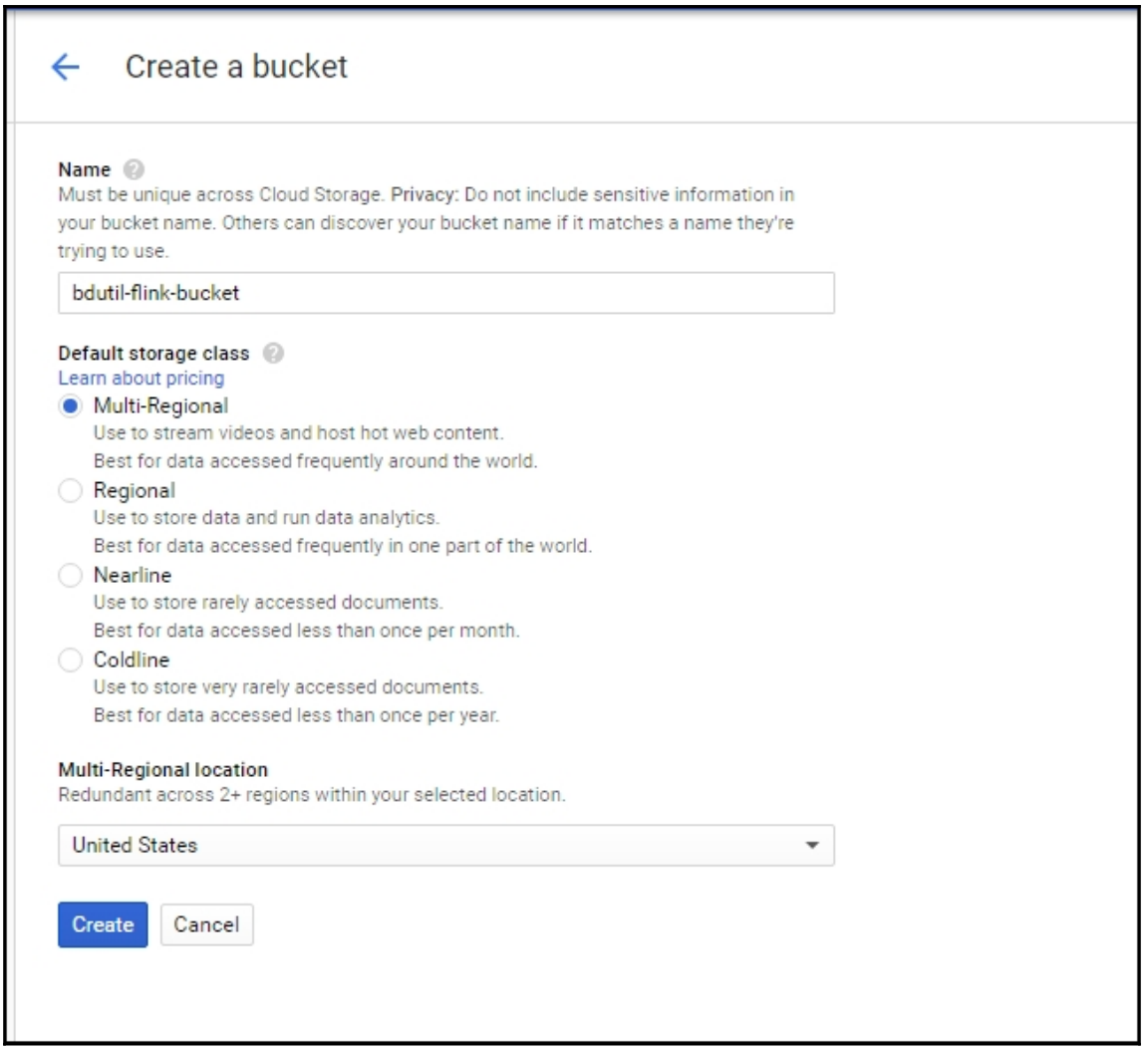

Services
Resource Groups
Tanmay Deshpande
Mumbai
Create Cluster - Quick Options
Go to advanced options
Click Here
General Configuration
Cluster name
My cluster
Logging
S3 folder
s3://aws-logs-458854225780-ap-south-1/elasticmapred
Launch mode
Cluster
Step execution
Software configuration
Vendor
Amazon
Release
emr-5.3.0
Applications
Core Hadoop: Hadoop 2.7.3 with Ganglia 3.7.2, Hive 2.1.1, Hue 3.11.0, Mahout 0.12.2, Pig 0.16.0, and Tez 0.8.4
HBase: HBase 1.2.3 with Ganglia 3.7.2, Hadoop 2.7.3, Hive 2.1.1, Hue 3.11.0, Phoenix 4.7.0, and ZooKeeper 3.4.9
Presto: Presto 0.152.3 with Hadoop 2.7.3 HDFS and Hive 2.1.1 Metastore
Spark: Spark 2.1.0 on Hadoop 2.7.3 YARN with Ganglia 3.7.2 and Zeppelin 0.6.2


Services
Resource Groups
Tanmay Deshpande
Mumbai
Create Cluster - Advanced Options
Go to quick options
Step 1: Software and Steps
Step 2: Hardware
Step 3: General Cluster Settings
Step 4: Security
Software Configuration
Vendor
Amazon
Release
emr-5.3.0
Hadoop 2.7.3
Flink 1.1.4
Pig 0.16.0
ZooKeeper 3.4.9
Hue 3.11.0
Spark 2.1.0
Zeppelin 0.6.2
Ganglia 3.7.2
Hive 2.1.1
Sqoop 1.4.6
Phoenix 4.7.0
HCatalog 2.1.1
Tez 0.8.4
HBase 1.2.3
Presto 0.157.1
Mahout 0.12.2
Oozie 4.3.0
Edit software settings (optional)
Enter configuration
Load JSON from S3
classification=config-file-name,properties=[myKey1=myValue1,myKey2=myValue2]
Add steps (optional)
Step type
Select a step
Configure

```
Welcome to the Google Cloud SDK!

To help improve the quality of this product, we collect anonymized usage data
 and anonymized stacktraces when crashes are encountered.. You may choose to opt
 out of this collection now (by choosing 'N' at the below prompt), or at any
 time in the future by running the following command:
    gcloud config set disable_usage_reporting true

Do you want to help improve the Google Cloud SDK (Y/n)?  Y


Your current Cloud SDK version is: 135.0.0
The latest available version is: 135.0.0
```

Components			
Status	Name	ID	Size
Not Installed	App Engine Go Extensions	app-engine-go	47.3 MiB
Not Installed	Cloud Datastore Emulator	cloud-datastore-emulator	15.4 MiB
Not Installed	Cloud Datastore Emulator (Legacy)	gcd-emulator	38.1 MiB
Not Installed	Cloud Pub/Sub Emulator	pubsub-emulator	16.3 MiB
Not Installed	Google Container Registry's Docker credential helper	docker-credential-gcr	2.2 MiB
Not Installed	gcloud Alpha Commands	alpha	< 1 MiB
Not Installed	gcloud Beta Commands	beta	< 1 MiB
Not Installed	gcloud app Java Extensions	app-engine-java	124.4 MiB
Not Installed	gcloud app Python Extensions	app-engine-python	7.2 MiB
Not Installed	kubectl	kubectl	15.9 MiB
Installed	BigQuery Command Line Tool	bq	< 1 MiB
Installed	Cloud SDK Core Libraries	core	5.1 MiB
Installed	Cloud Storage Command Line Tool	gsutil	2.8 MiB
Installed	Default set of gcloud commands	gcloud	

```
To install or remove components at your current SDK version [135.0.0], run:
  $ gcloud components install COMPONENT_ID
  $ gcloud components remove COMPONENT_ID

To update your SDK installation to the latest version [135.0.0], run:
  $ gcloud components update


Modify profile to update your $PATH and enable shell command
completion? (Y/n)?
```


Create a bucket
Name
Must be unique across Cloud Storage. Privacy: Do not include sensitive information in your bucket name. Others can discover your bucket name if it matches a name they're trying to use.
bdutil-flink-bucket
Default storage class
Learn about pricing
Multi-Regional
Use to stream videos and host hot web content.
Best for data accessed frequently around the world.
Regional
Use to store data and run data analytics.
Best for data accessed frequently in one part of the world.
Nearline
Use to store rarely accessed documents.
Best for data accessed less than once per month.
Coldline
Use to store very rarely accessed documents.
Best for data accessed less than once per year.
Multi-Regional location
Redundant across 2+ regions within your selected location.
United States
Create
Cancel

```
[tdeshpande2@dev-instance-1 bdutil-master]$ sudo ./bdutil -e extensions/flink/flink_env.sh deploy
Sat Nov 19 05:01:18 UTC 2016: Using local tmp dir for staging files: /tmp/bdutil-20161119-050118-FpC
Sat Nov 19 05:01:18 UTC 2016: Using custom environment-variable file(s): bdutil_env.sh extensions/flink/flink_env.sh
Sat Nov 19 05:01:18 UTC 2016: Reading environment-variable file: ./bdutil_env.sh
Sat Nov 19 05:01:18 UTC 2016: Reading environment-variable file: extensions/flink/flink_env.sh
Sat Nov 19 05:01:18 UTC 2016: No explicit GCE_MASTER_MACHINE_TYPE provided; defaulting to value of GCE_MACHINE_TYPE: n1-standar
Deploy cluster with following settings?
      CONFIGBUCKET='bdutil-flink-bucket'
      PROJECT='
      GCE_IMAGE='https://www.googleapis.com/compute/v1/projects/debian-cloud/global/images/backports-debian-7-wheezy-v20160531'
      GCE_ZONE='europe-west1-d'
      GCE_NETWORK='default'
      GCE_TAGS='bdutil'
      PREEMPTIBLE_FRACTION=0.0
      PREFIX='hadoop'
      NUM_WORKERS=2
      MASTER_HOSTNAME='hadoop-m'
      WORKERS='hadoop-w-0 hadoop-w-1'
      BDUTIL_GCS_STAGING_DIR='gs://bdutil-flink-bucket/bdutil-staging/hadoop-m'
      (y/n)
```