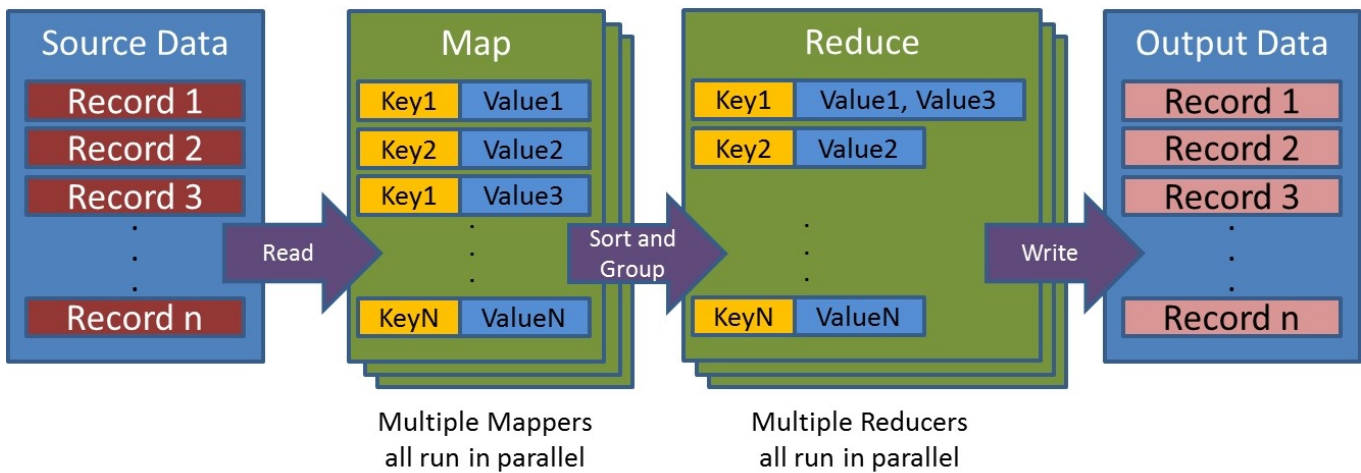
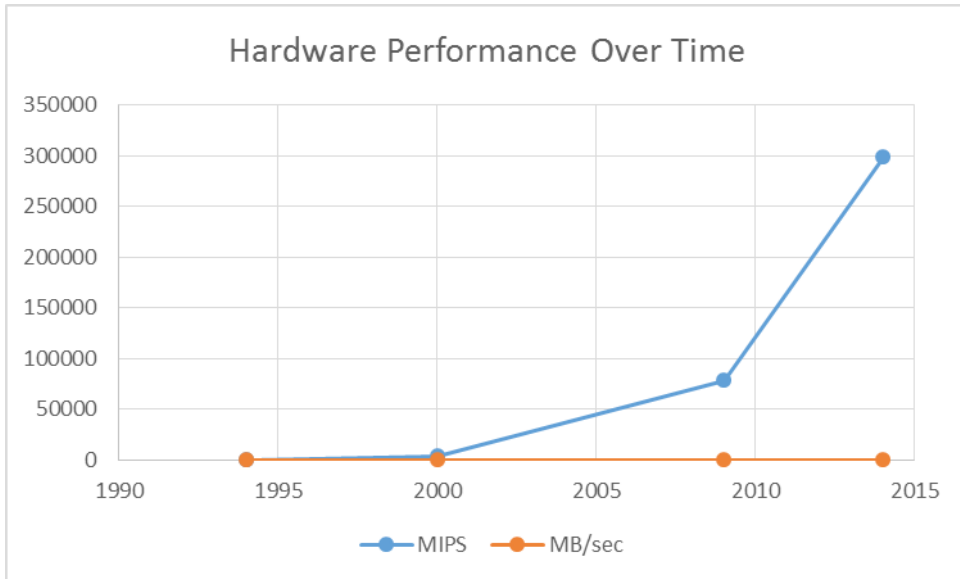
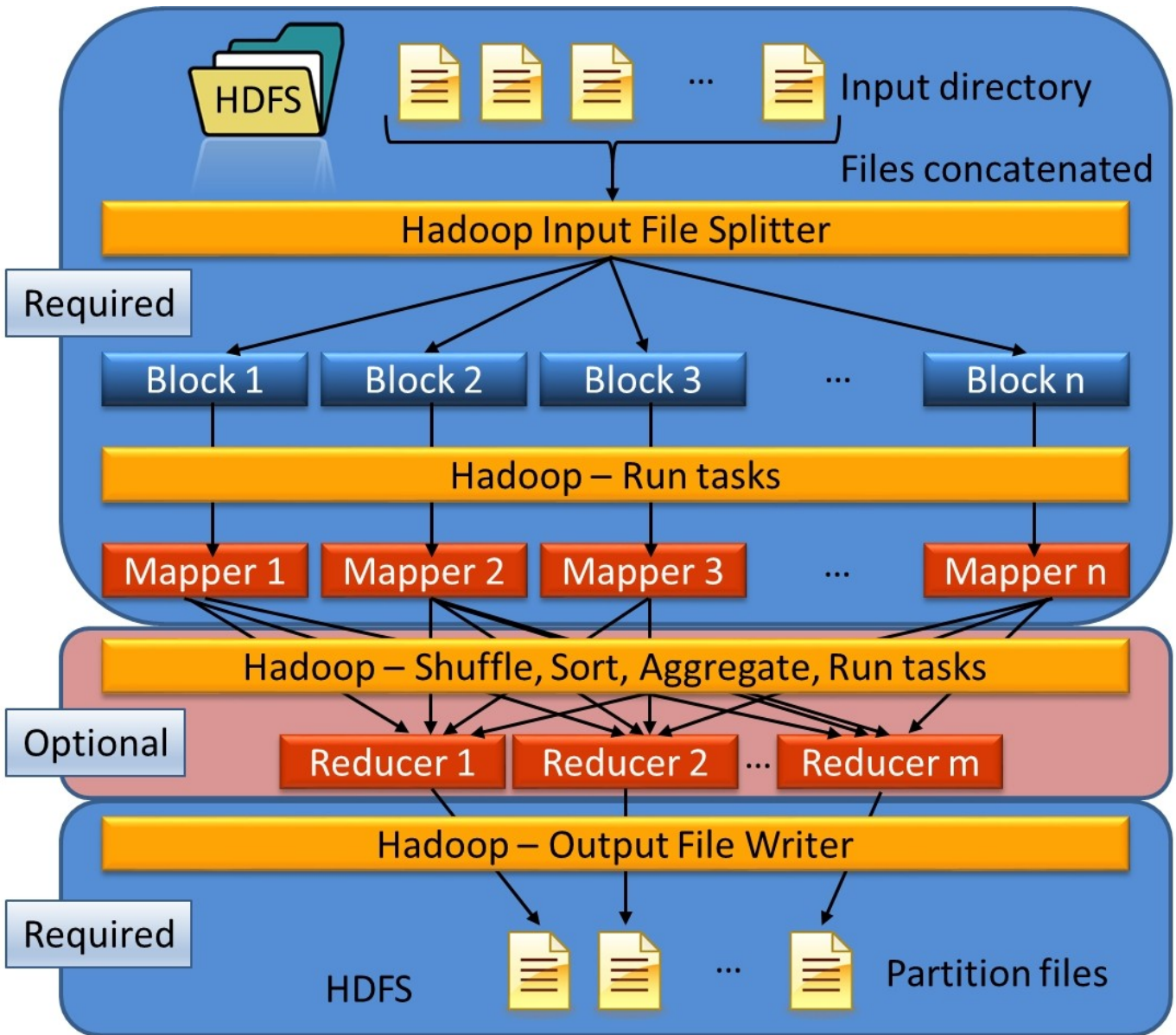
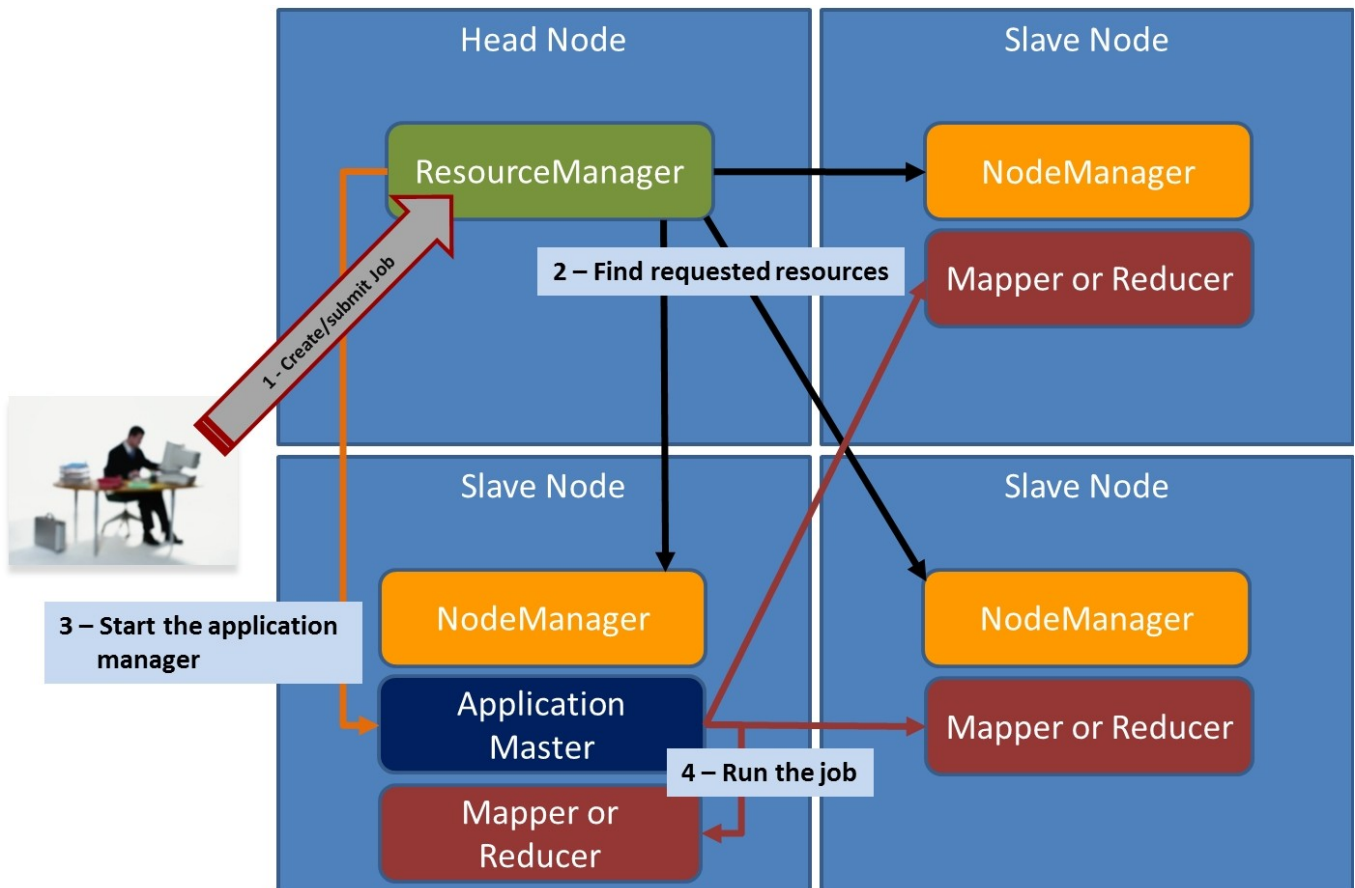
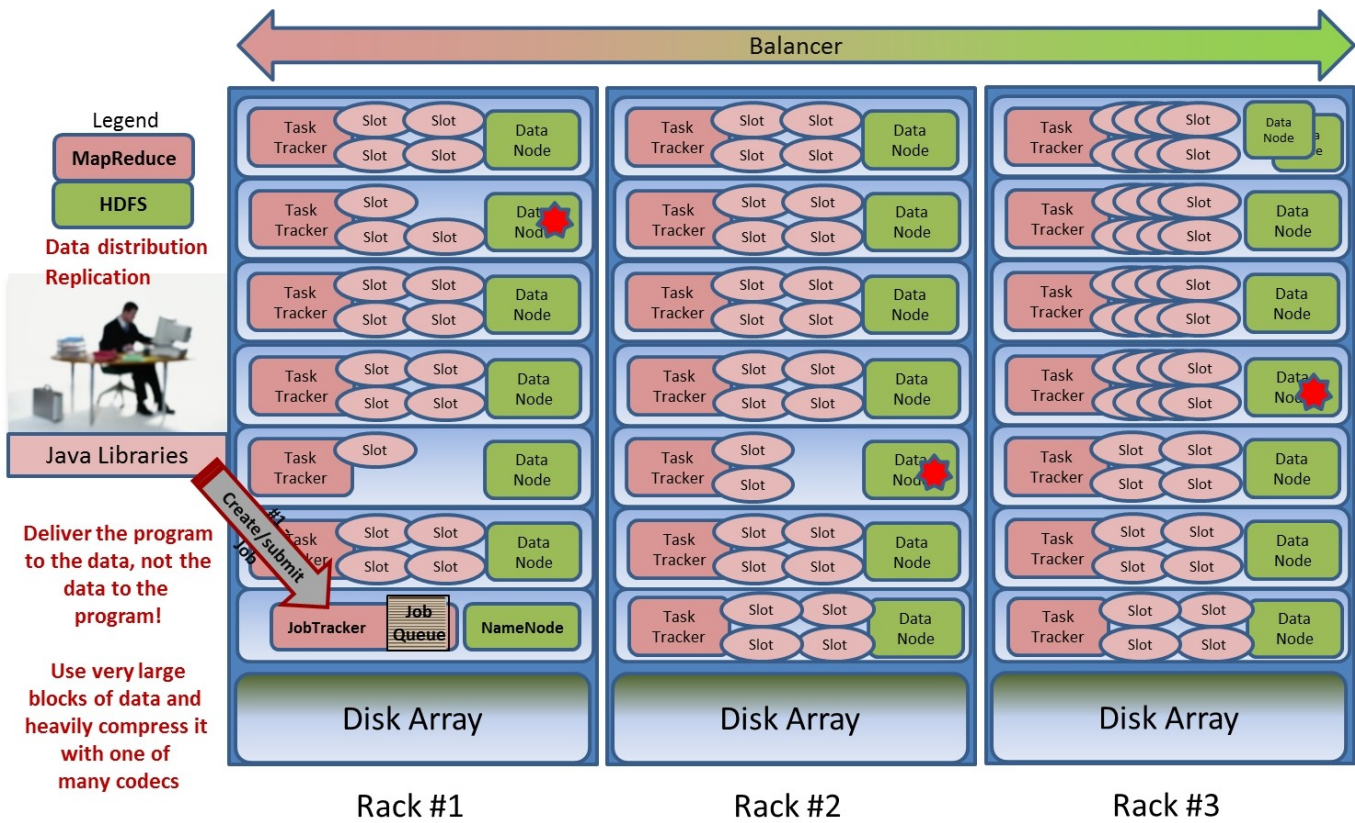


Learning Cascading

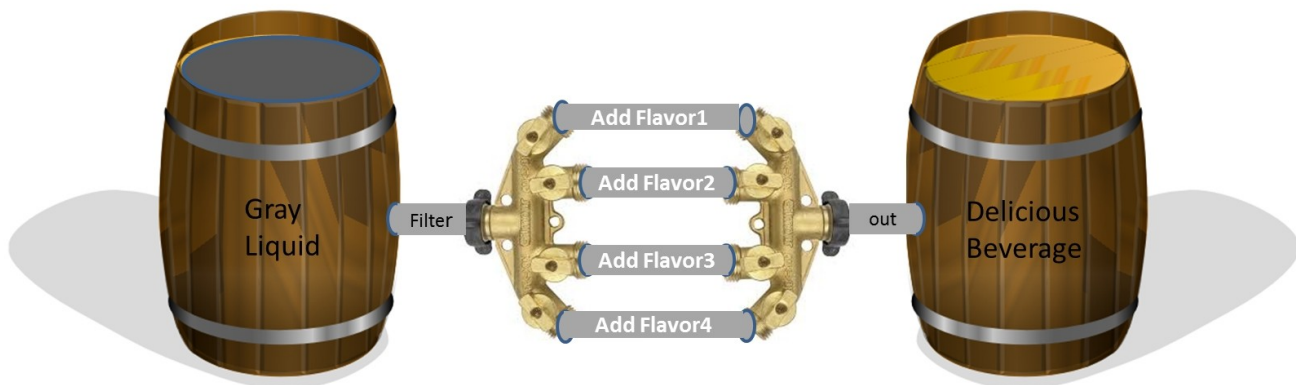
Chapter 1: The Big Data Core Technology Stack

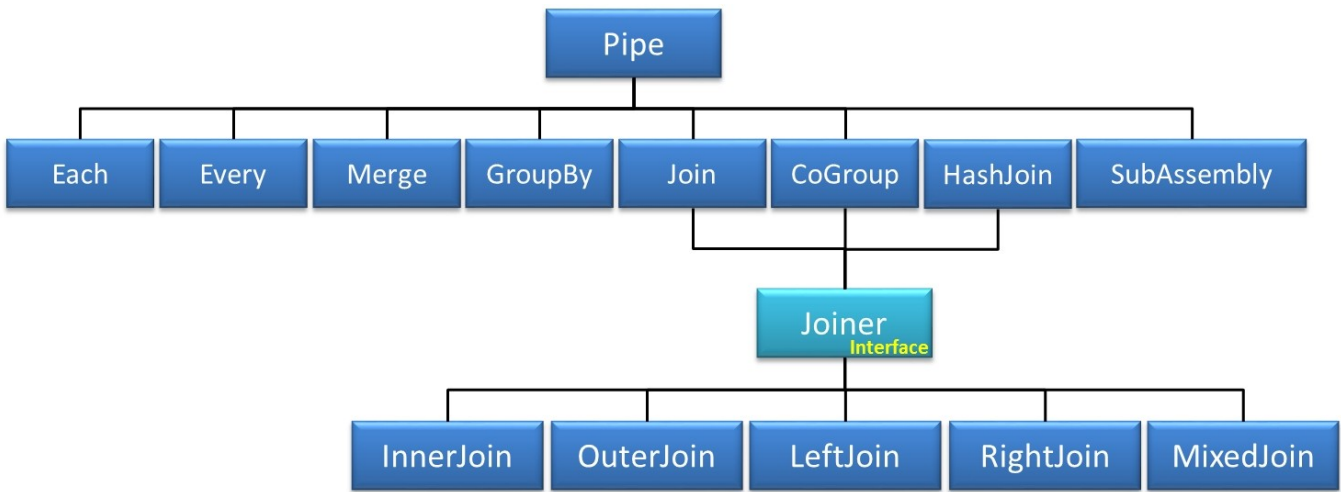






Chapter 2: Cascading Basics in Detail





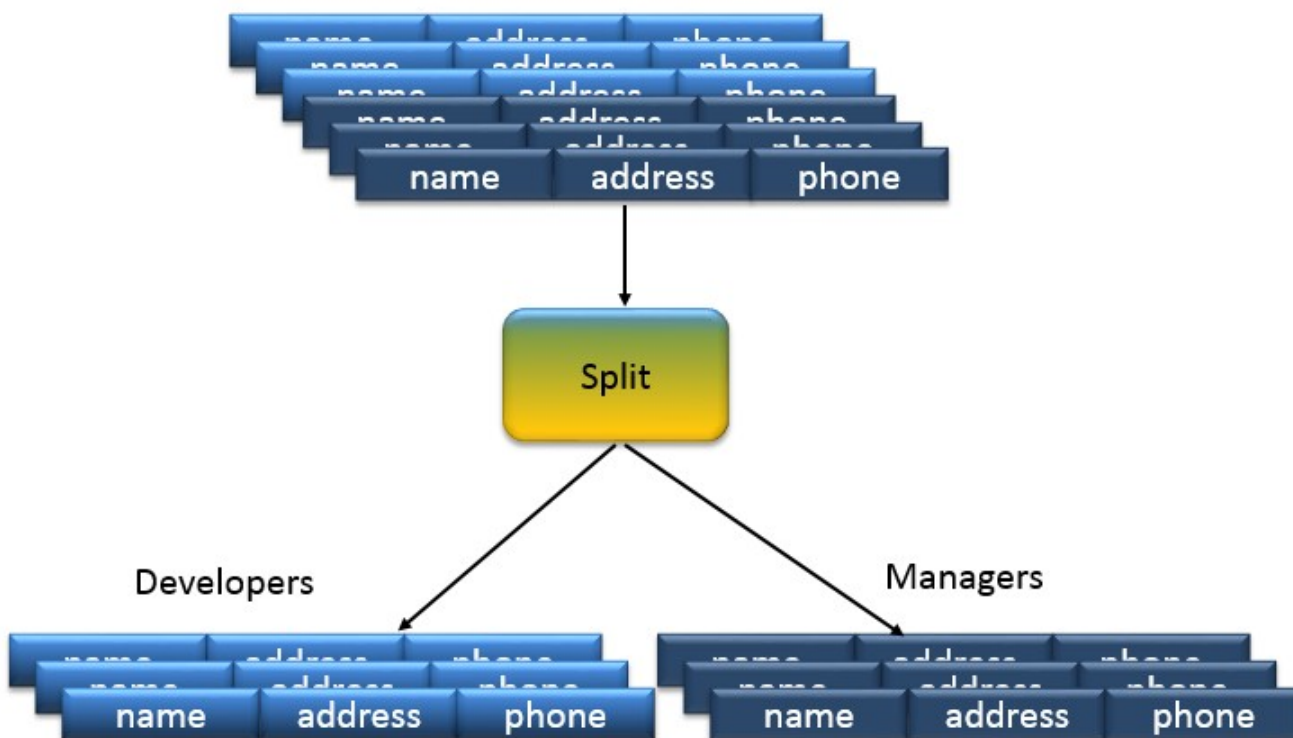
Payroll Data

name	division	salary
------	----------	--------

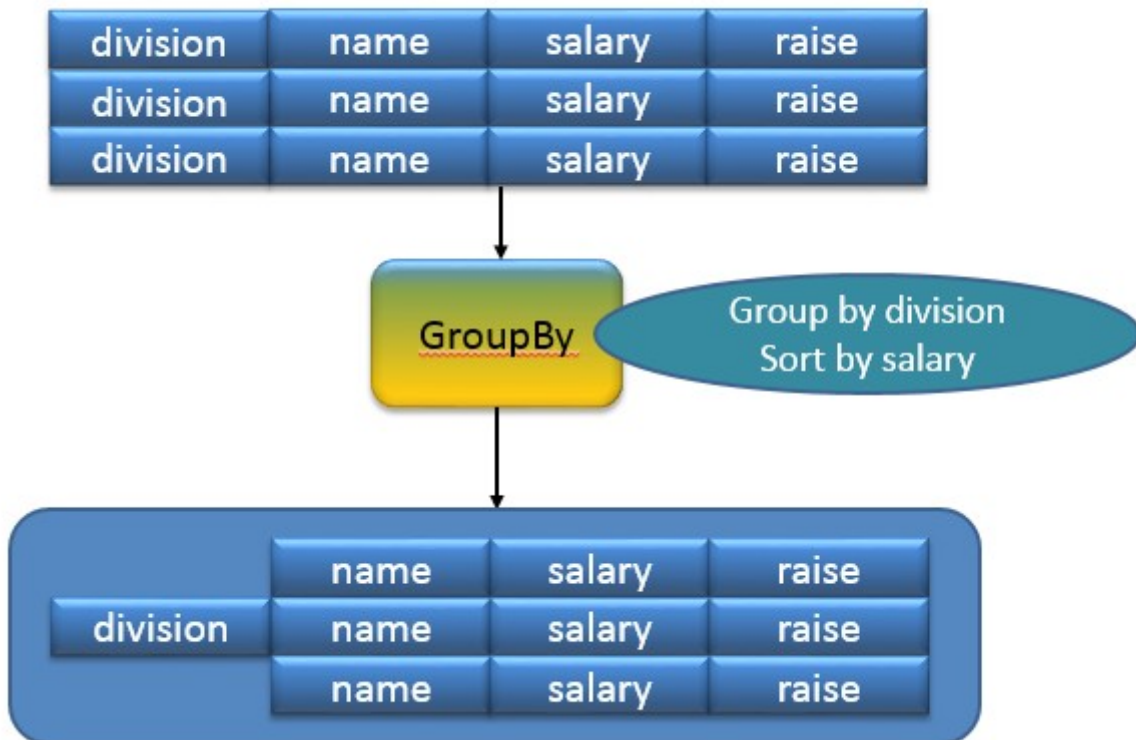


name	division	salary	raise
------	----------	--------	-------

HR Data



Payroll Data



Payroll Data

	name	salary	raise
division	name	salary	raise
	name	salary	raise



division	<u>tot_salary</u>	<u>tot_raise</u>
----------	-------------------	------------------

HR Data

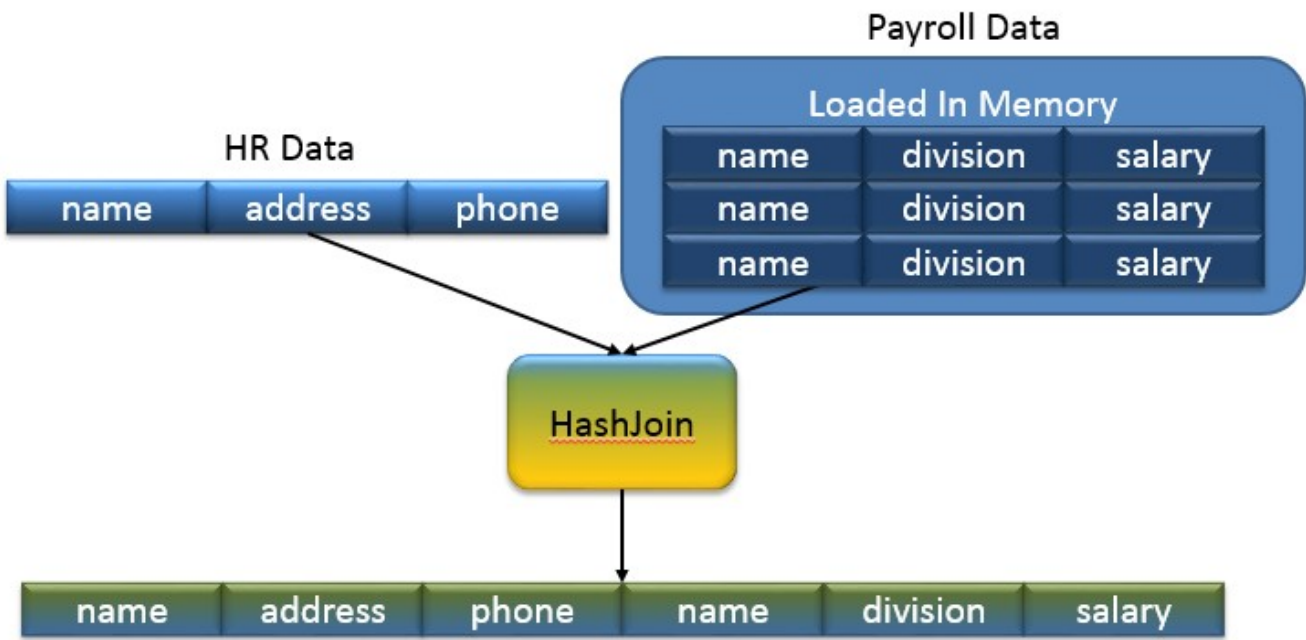
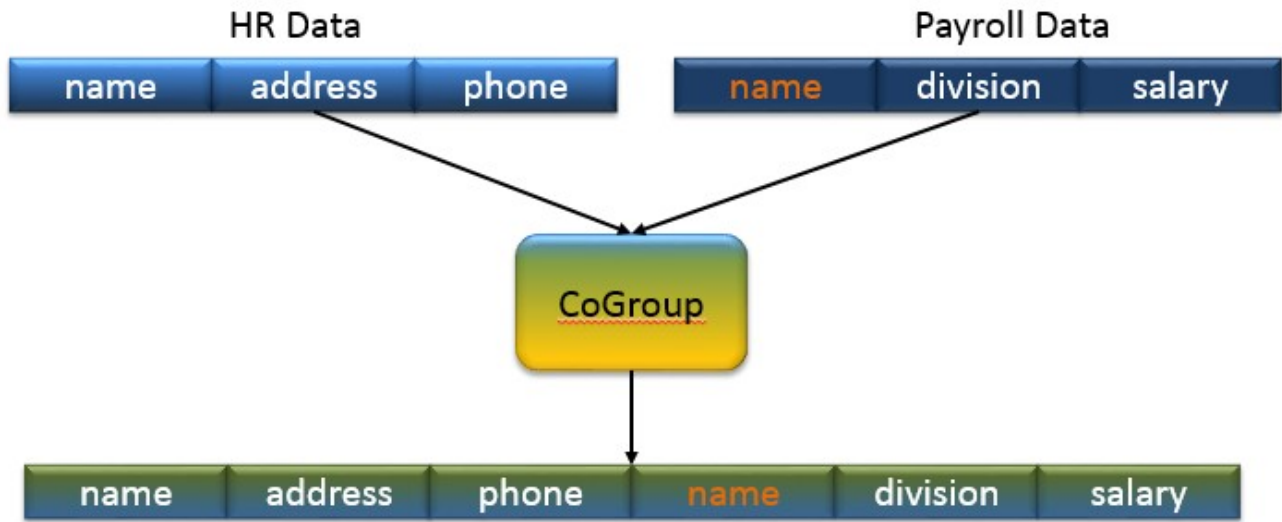
name	address	phone
name	address	phone
name	address	phone

HR Data Update

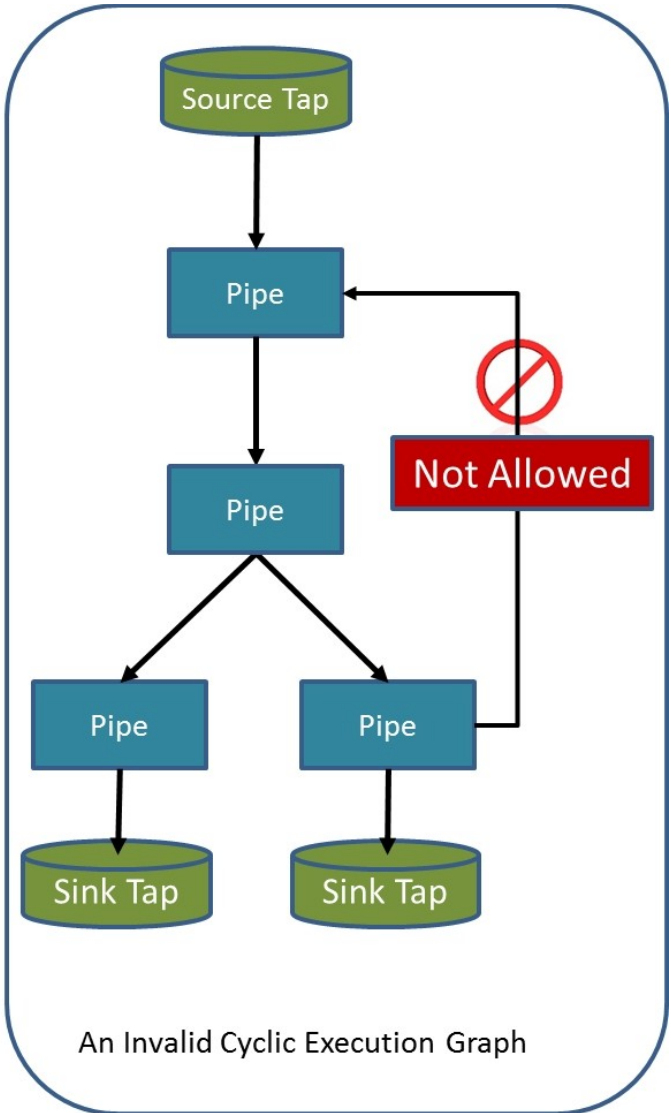
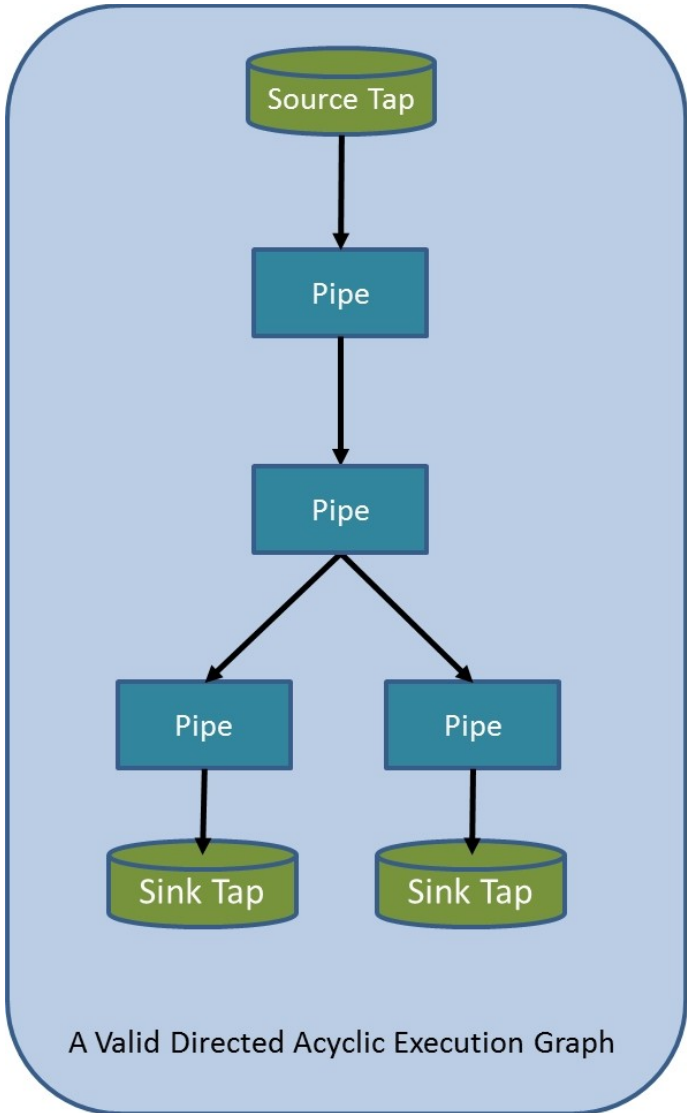
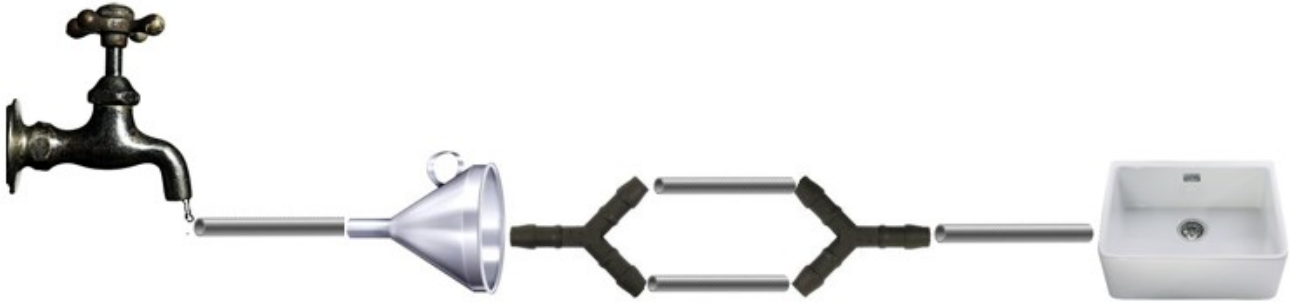
name	address	phone
name	address	phone
name	address	phone

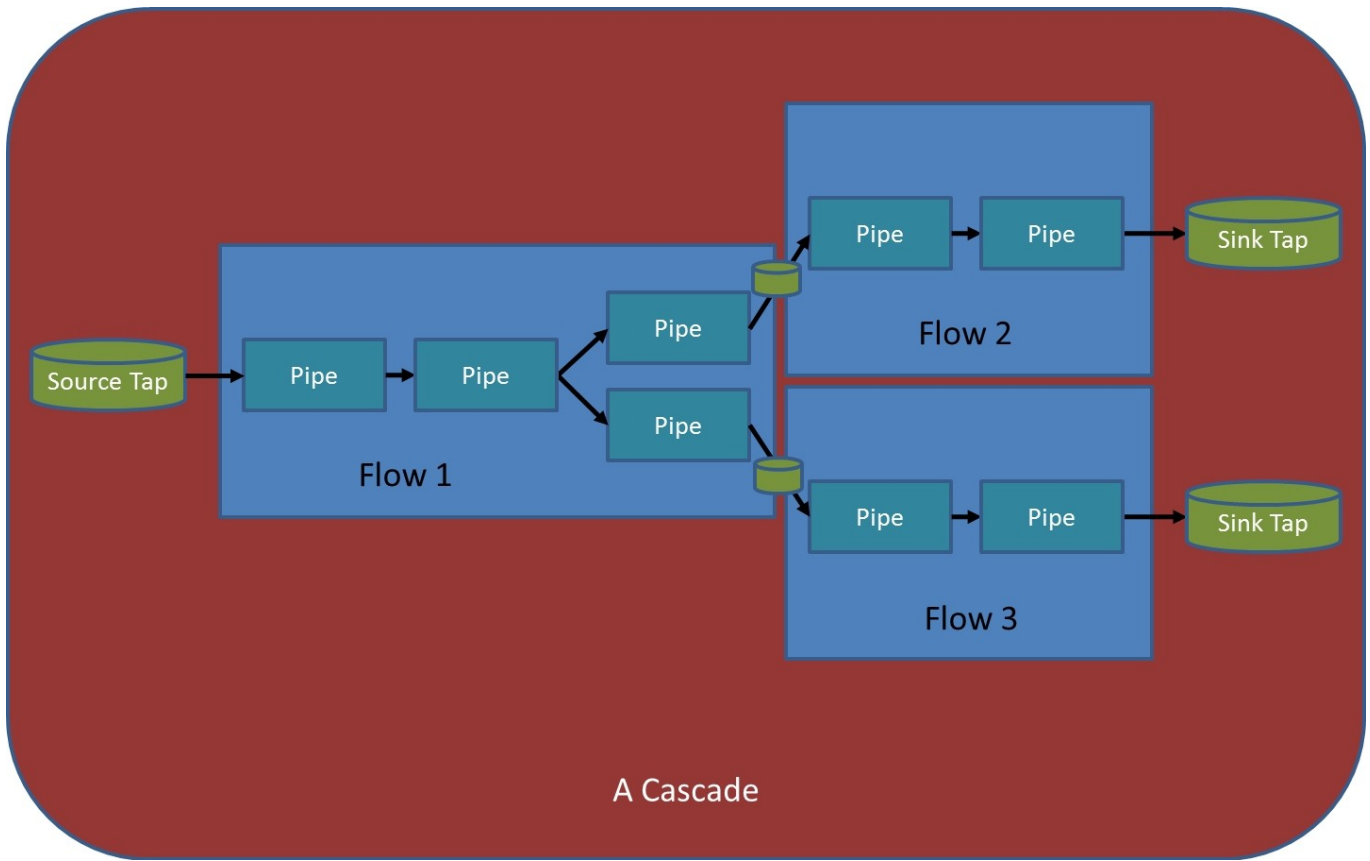


name	address	phone
name	address	phone
name	address	phone
name	address	phone
name	address	phone

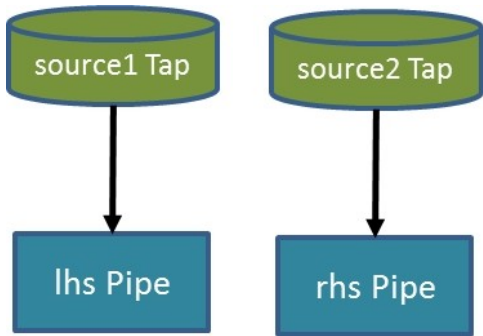


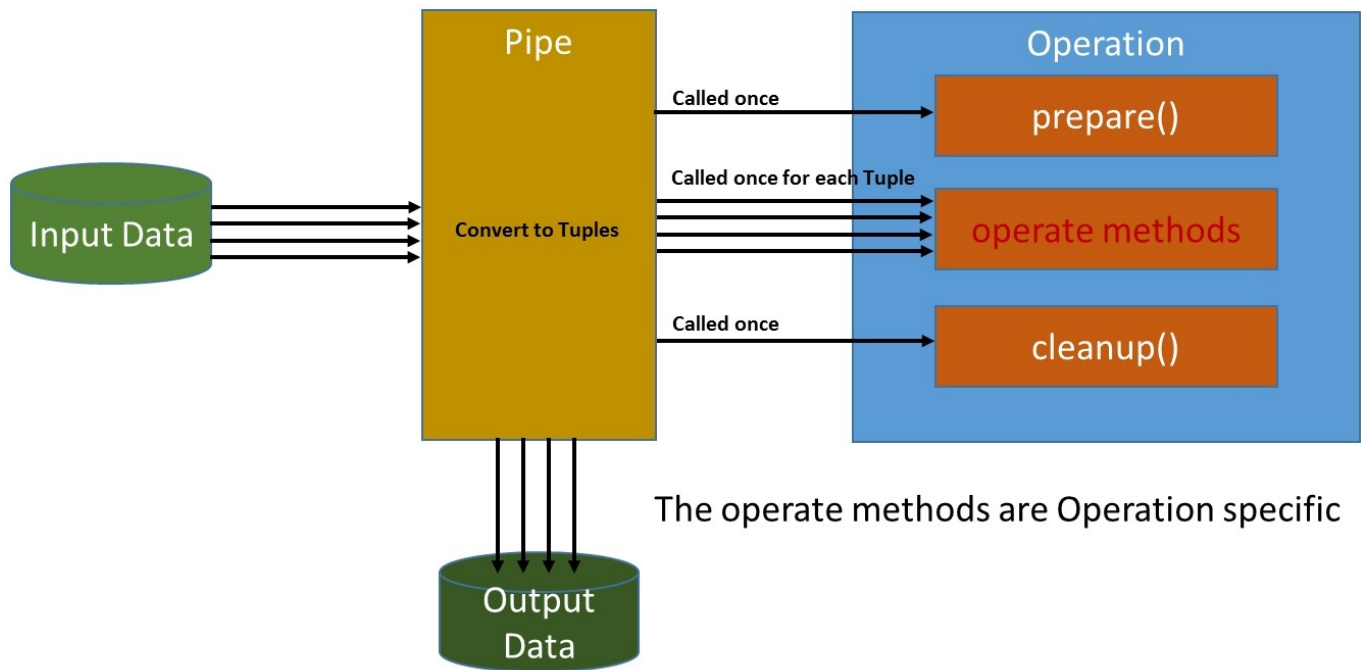
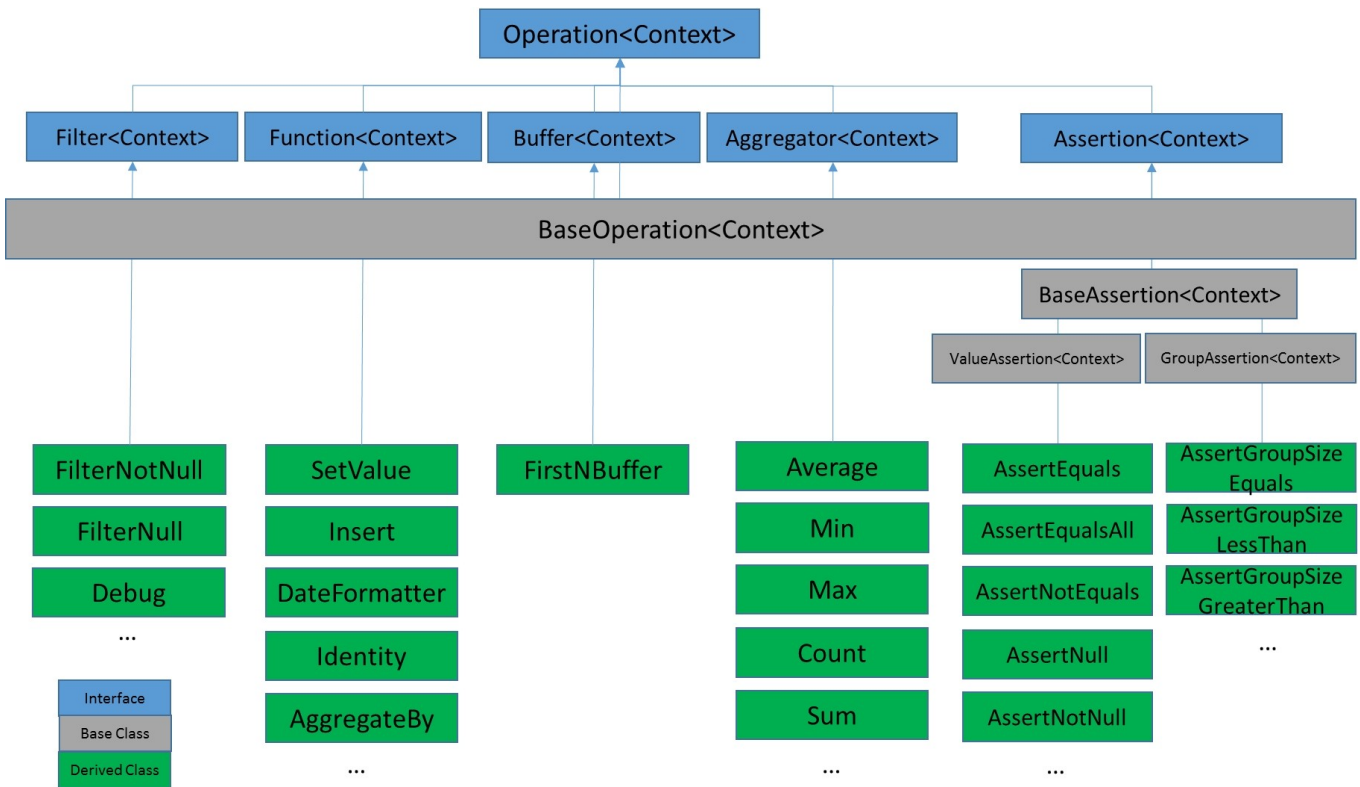
	Default	Filter	Function	Aggregator	Buffer
Each	-	ALL	RESULTS	-	-
Every	-	ALL	-	GROUP + RESULTS	ALL
CoGroup	ALL	-	-	-	-
GroupBy	ALL	-	-	-	-
Merge	ALL	-	-	-	-
HashJoin	ALL	-	-	-	-



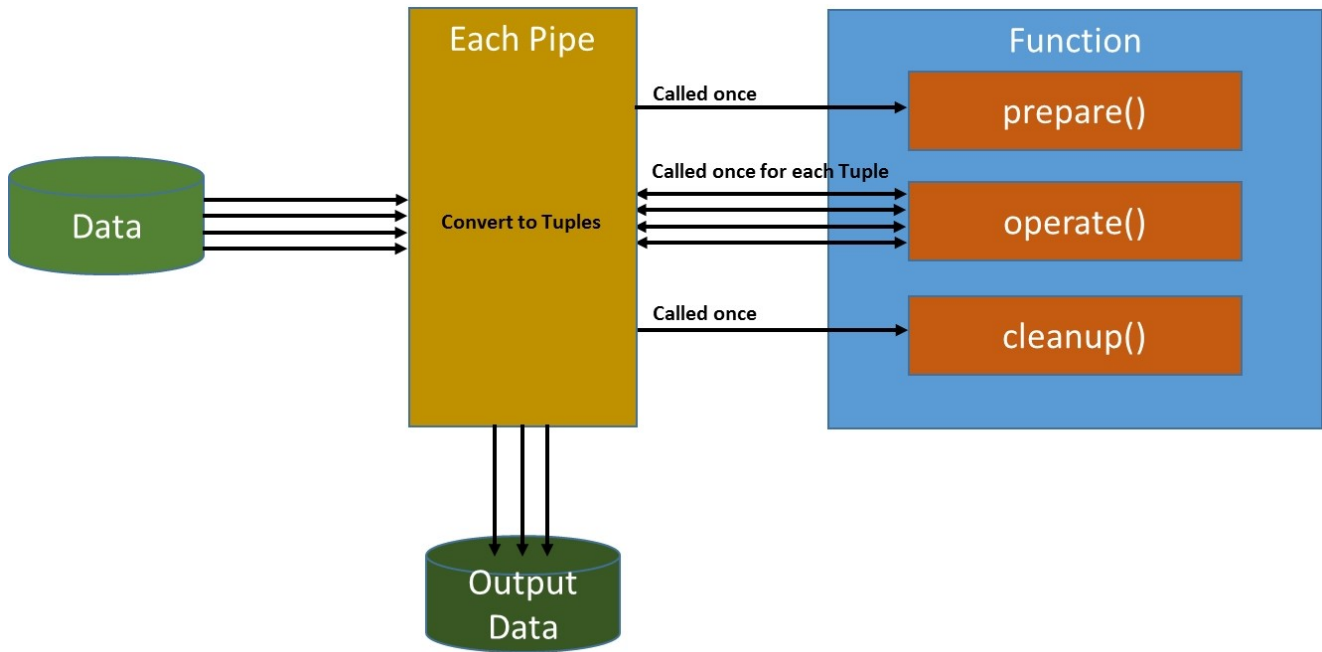
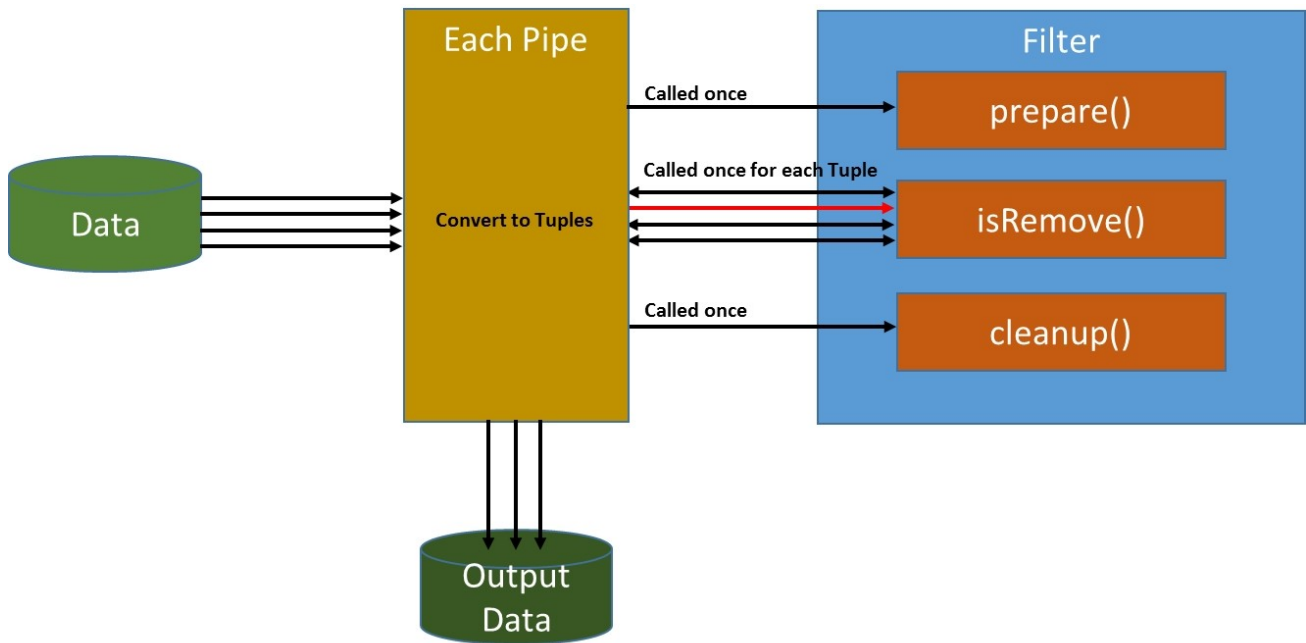


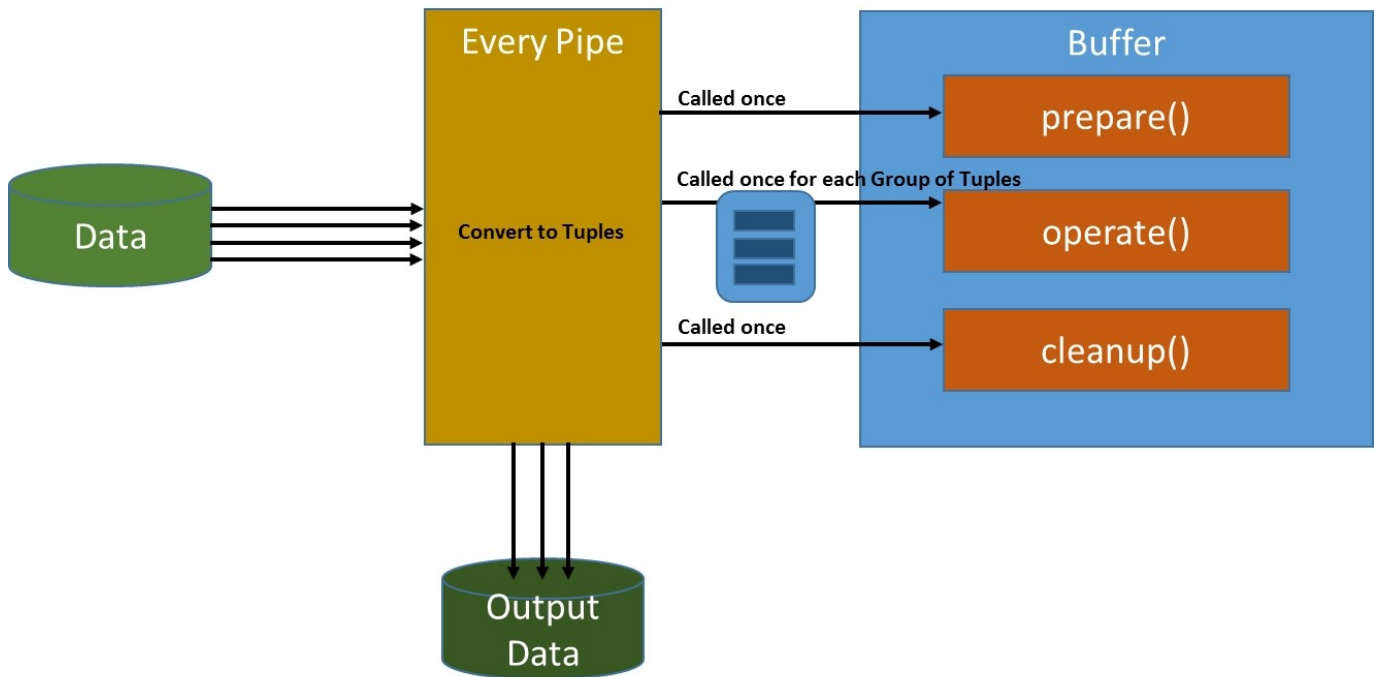
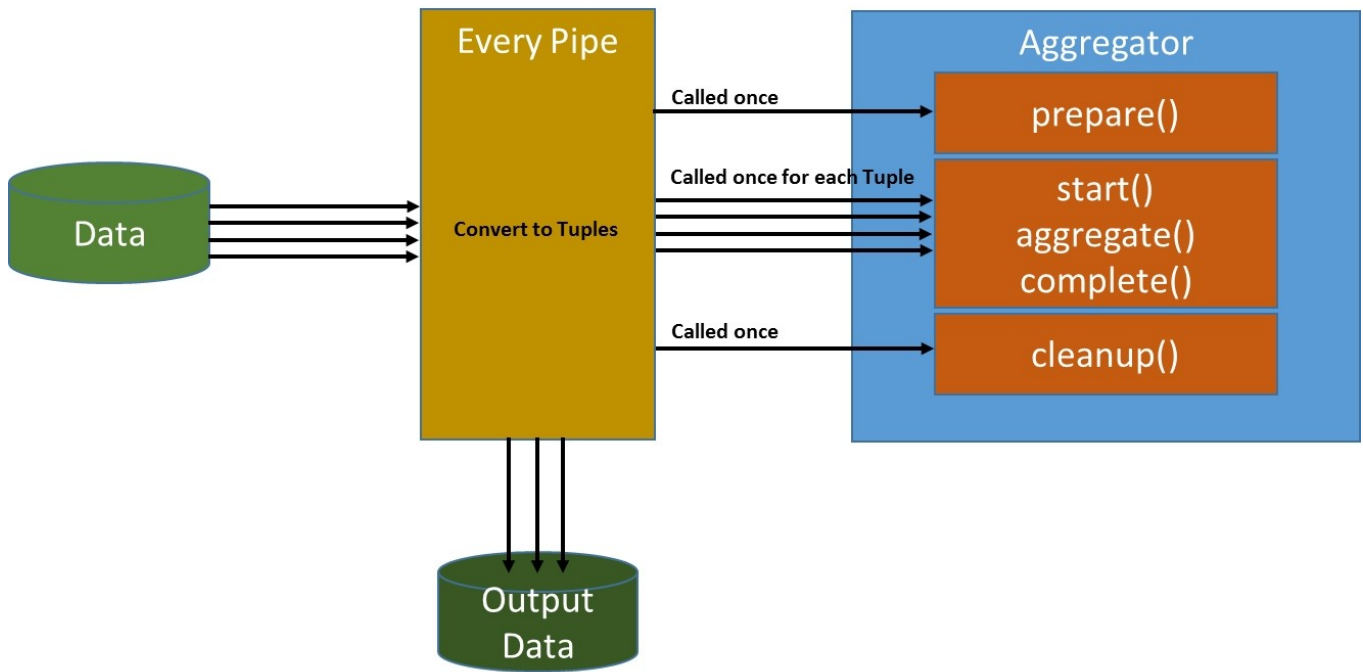
Chapter 3: Understanding Custom Operations

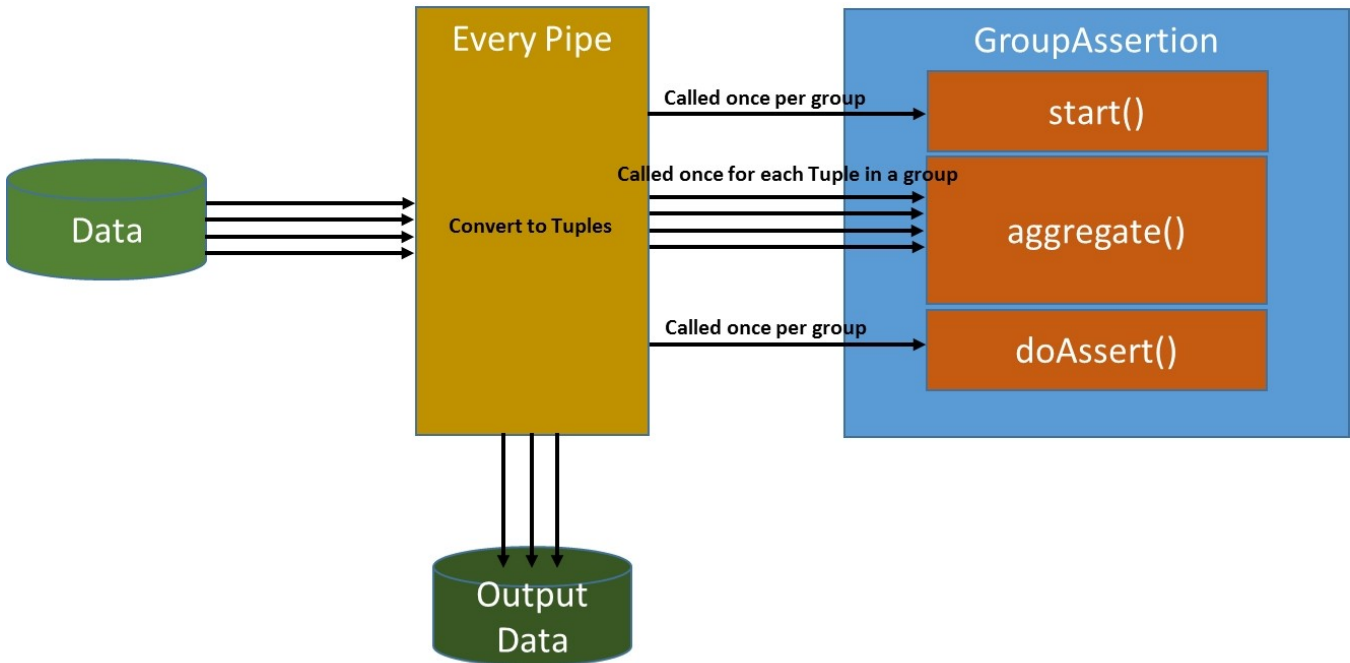
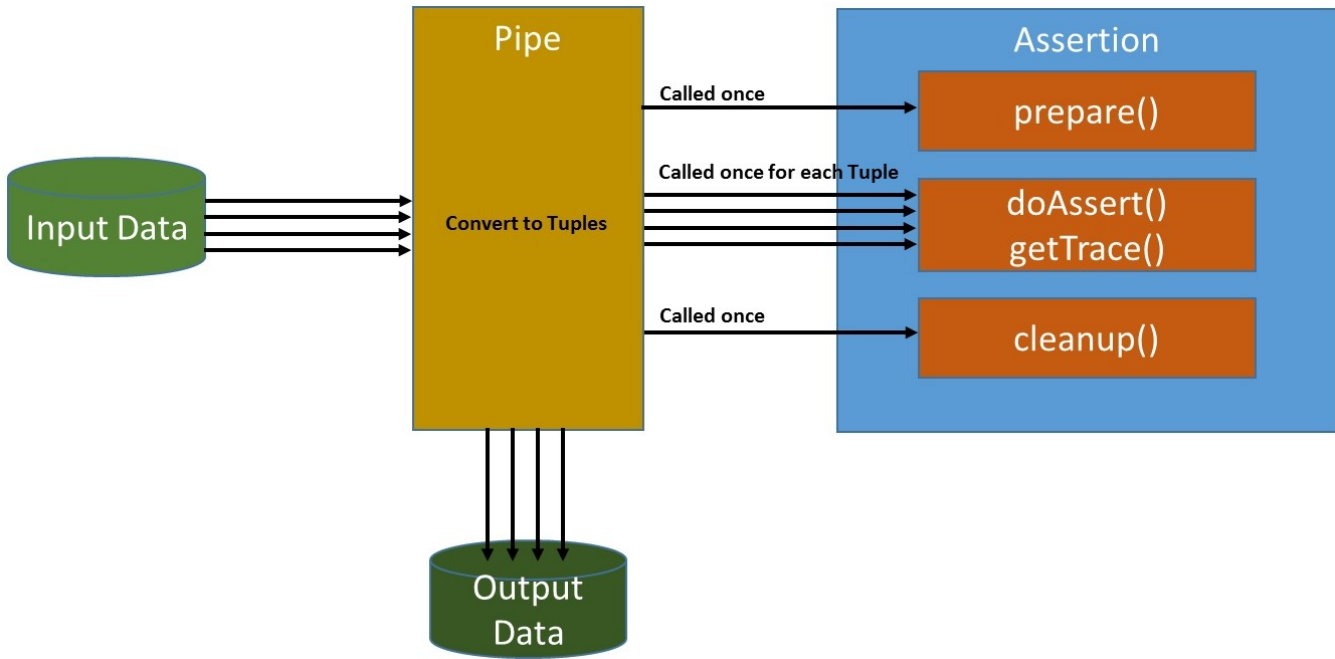


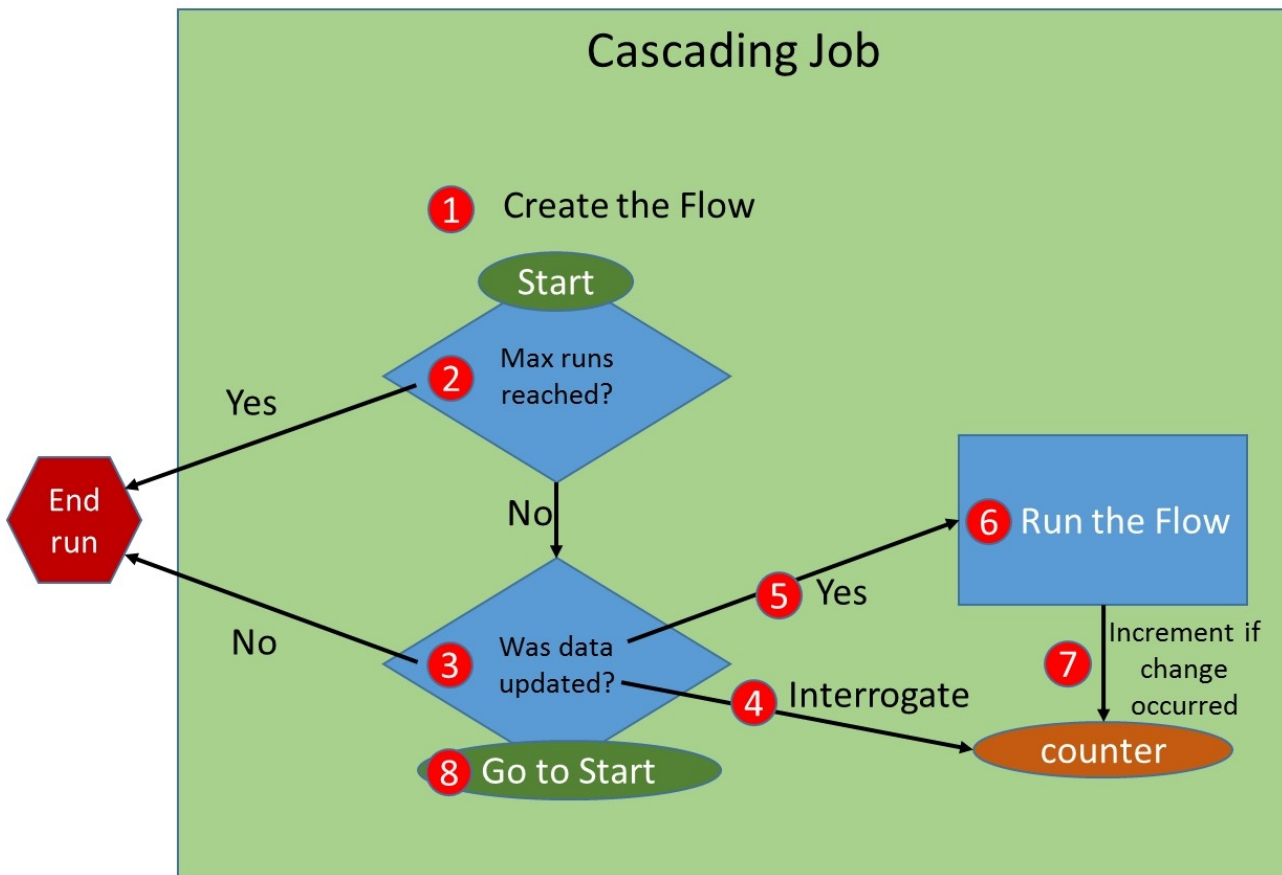
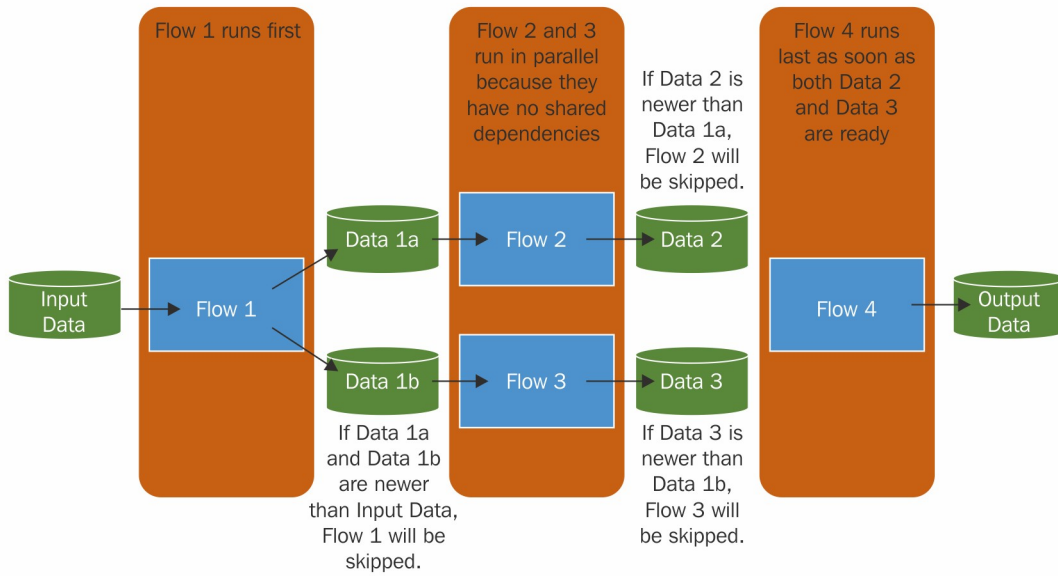


Interface	getContext()	setContext()	getArguments()	getDeclaredFields()	getOutputCollector()	getGroup()	getArgumentsIterator()
OperationCall	X	X					
FunctionCall			X	X	X		
FilterCall			X				
AggregatorCall			X	X	X	X	
BufferCall			X	X	X	X	X
ValueAssertionCall			X				
GroupAssertionCall			X			X	

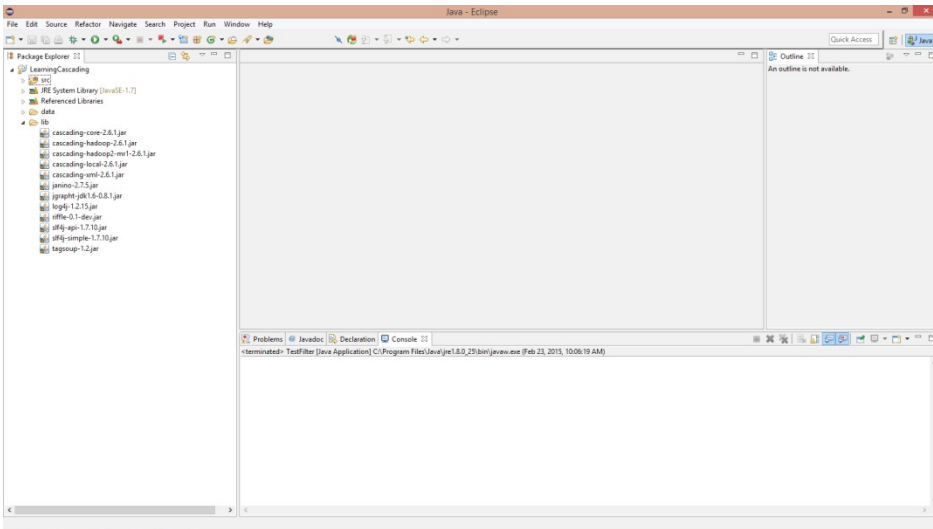
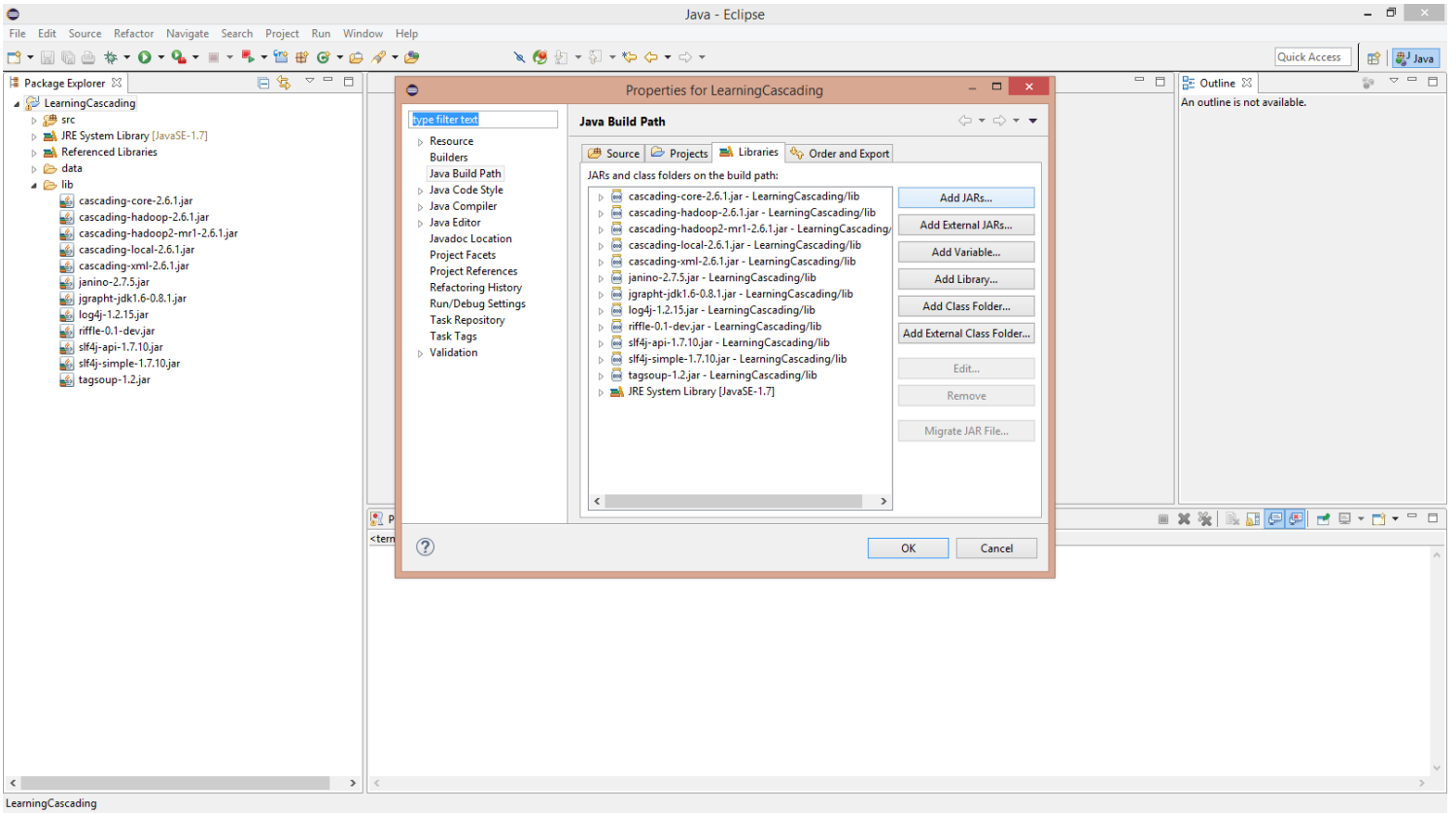


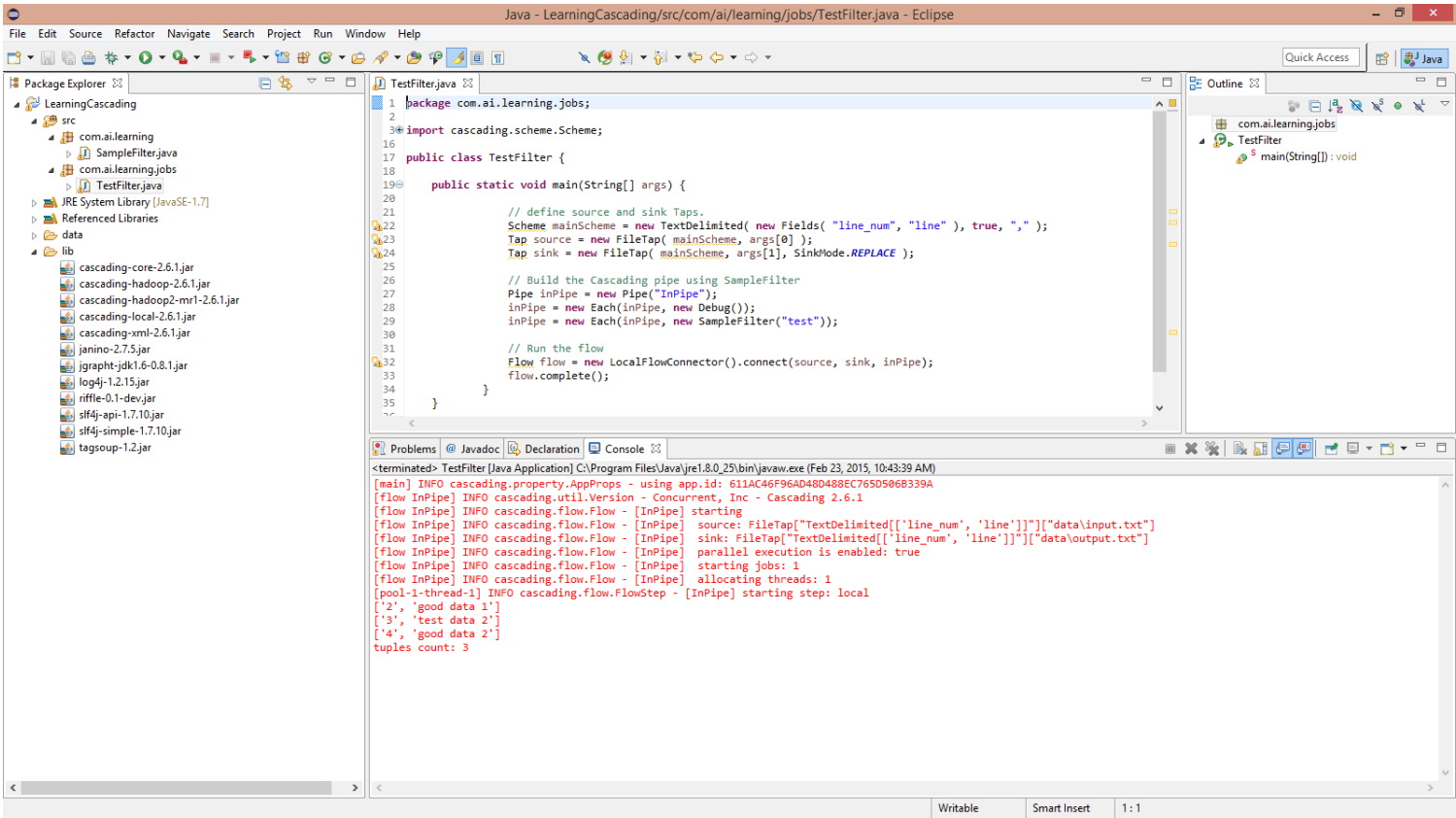
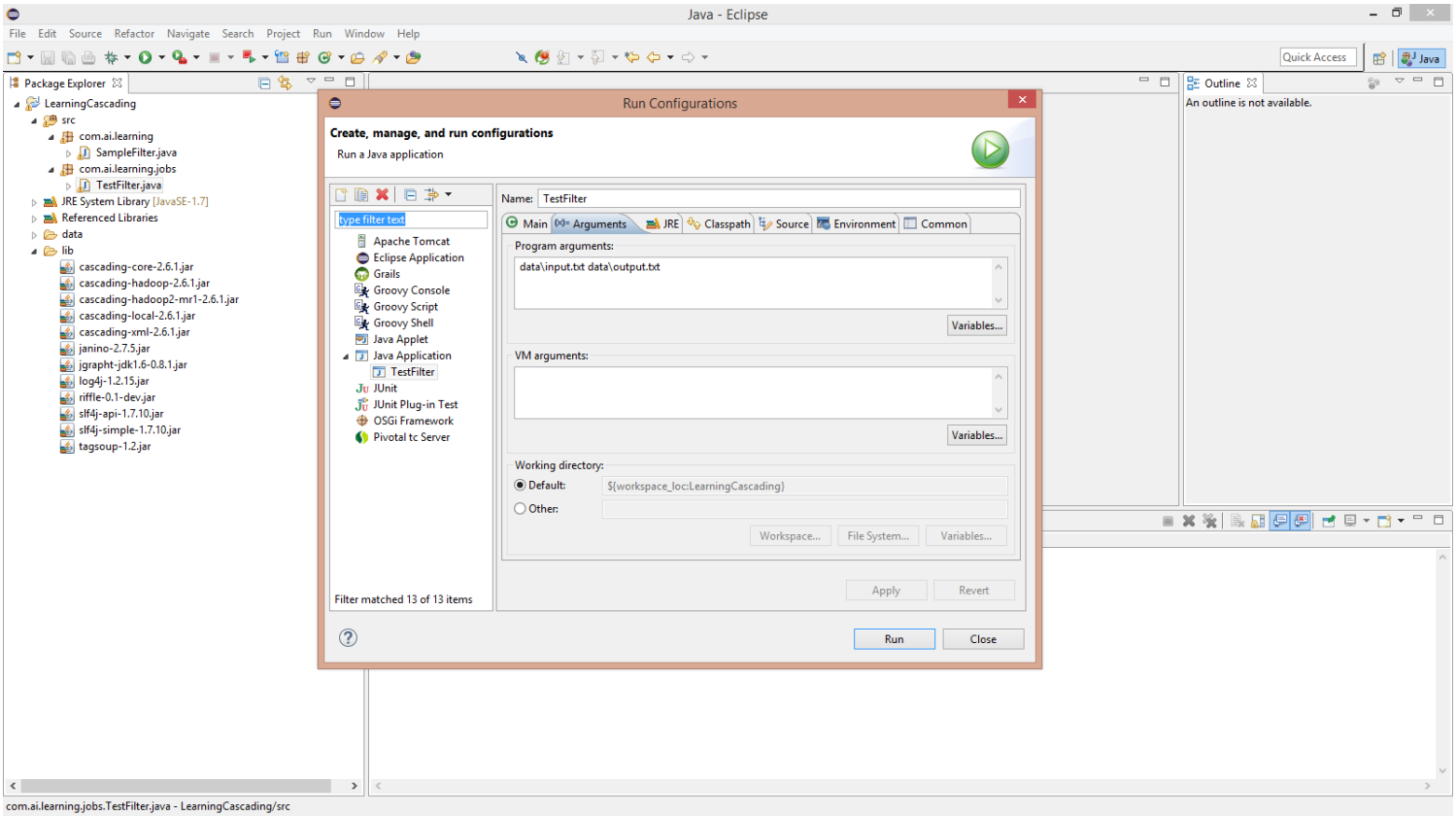


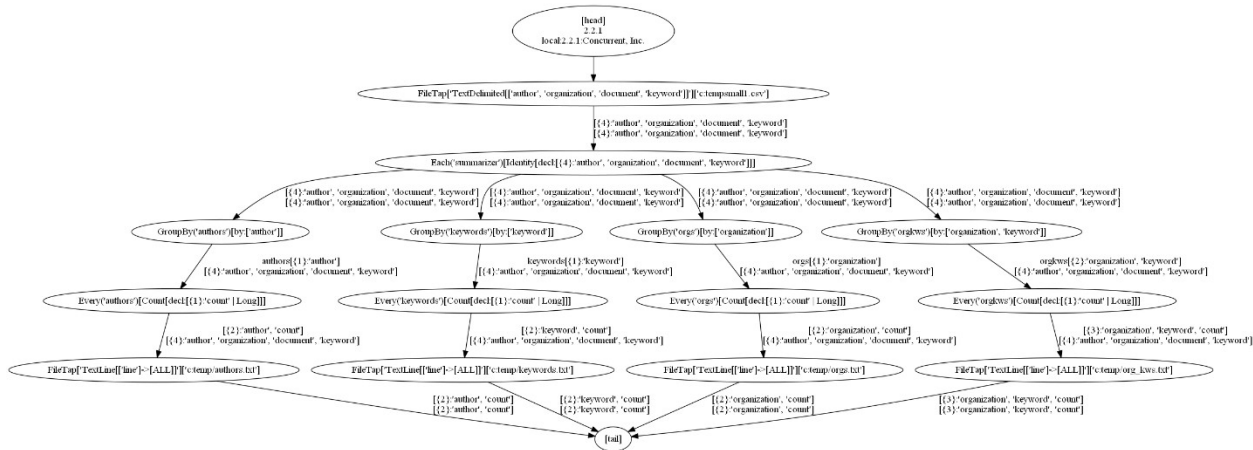
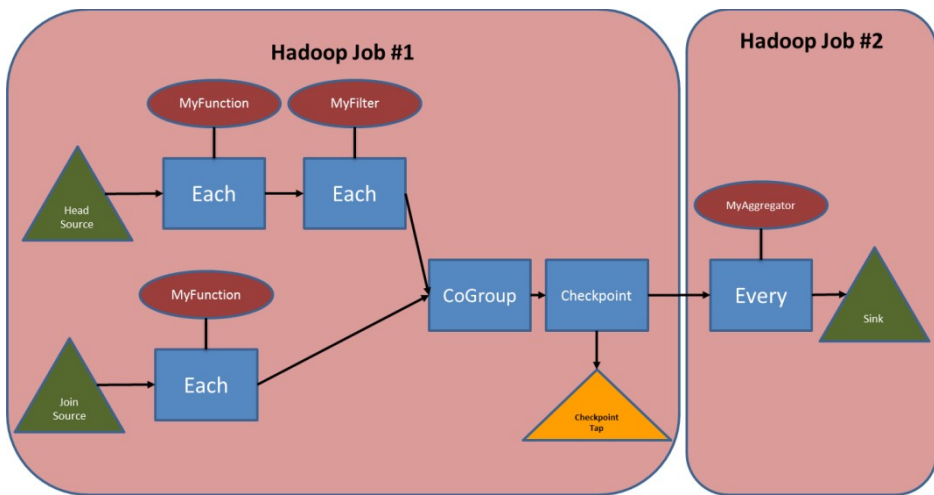
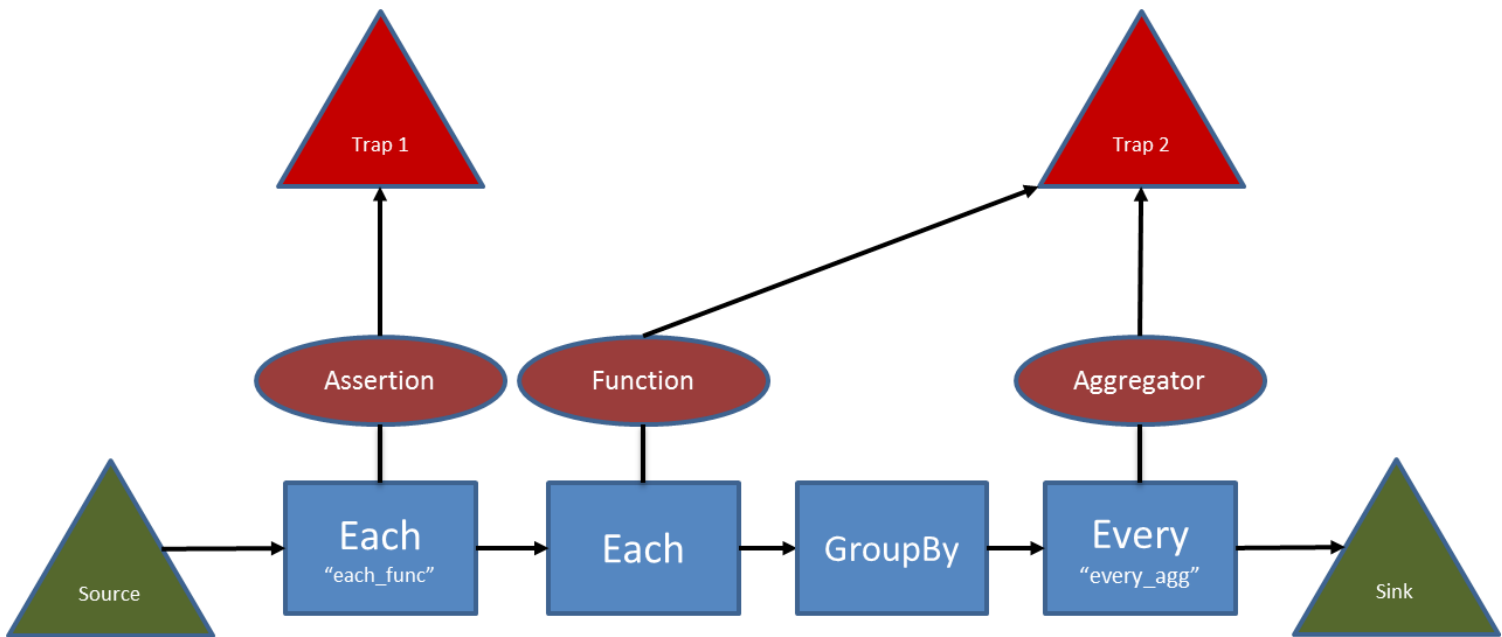




Chapter 6: Testing a Cascading Application





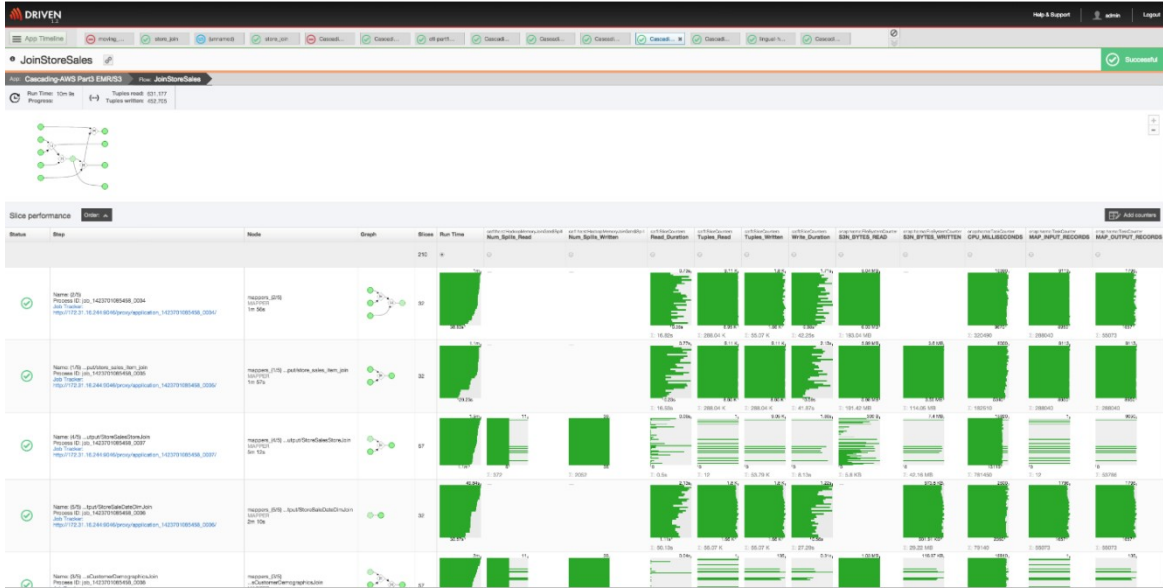
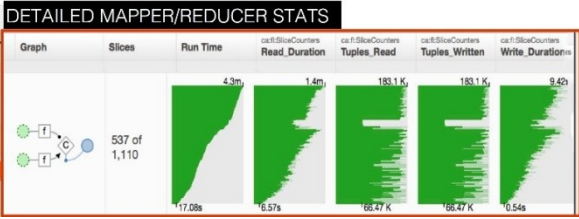
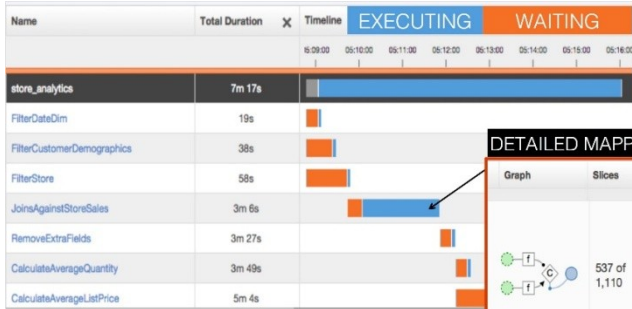


[(1/4) authors]
 src:[small]
 gp:authors
 snk:[authors]

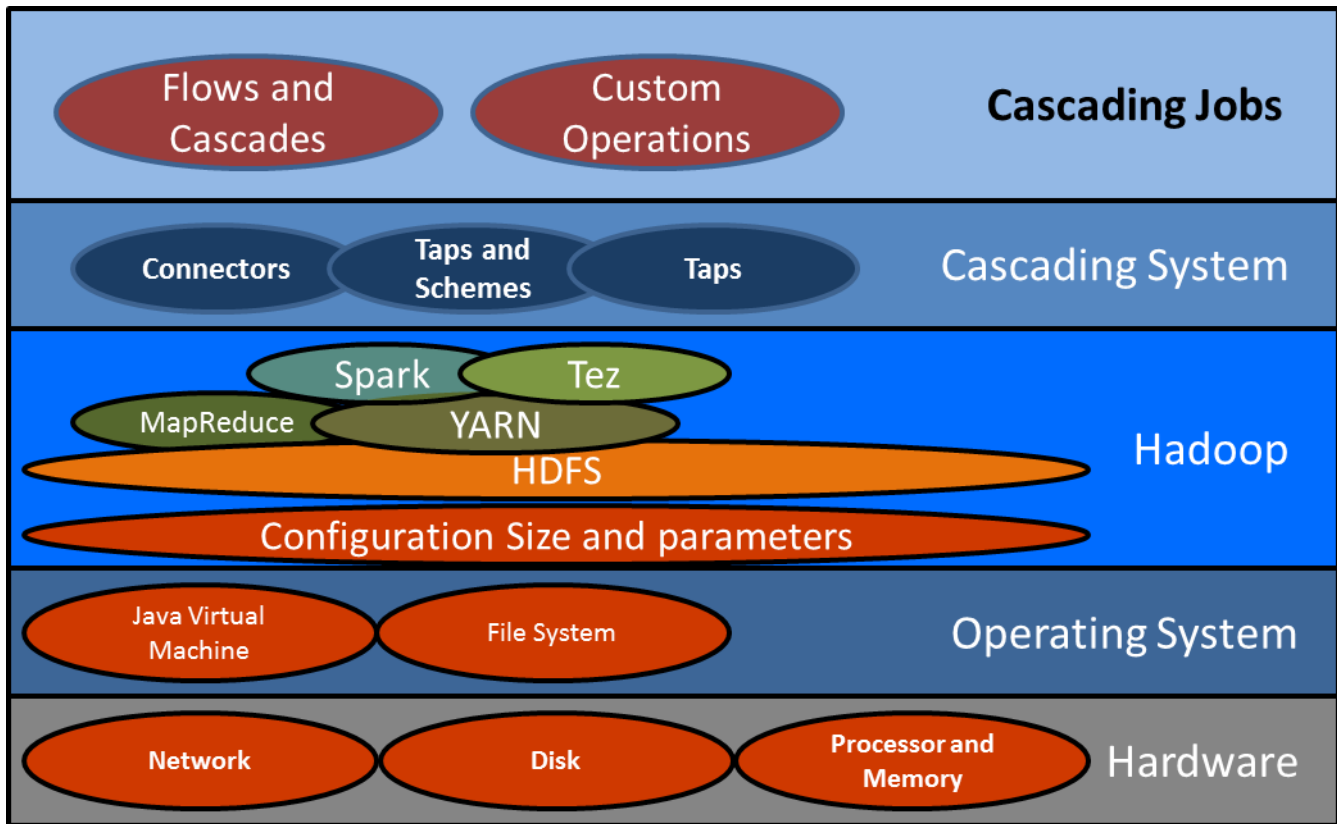
[(2/4) orgs]
 src:[small]
 gp:orgs
 snk:[orgs]

[(3/4) org_kws]
 src:[small]
 gp:orgkws
 snk:[org_kws]

[(4/4) keywords]
 src:[small]
 gp:keywords
 snk:[keywords]

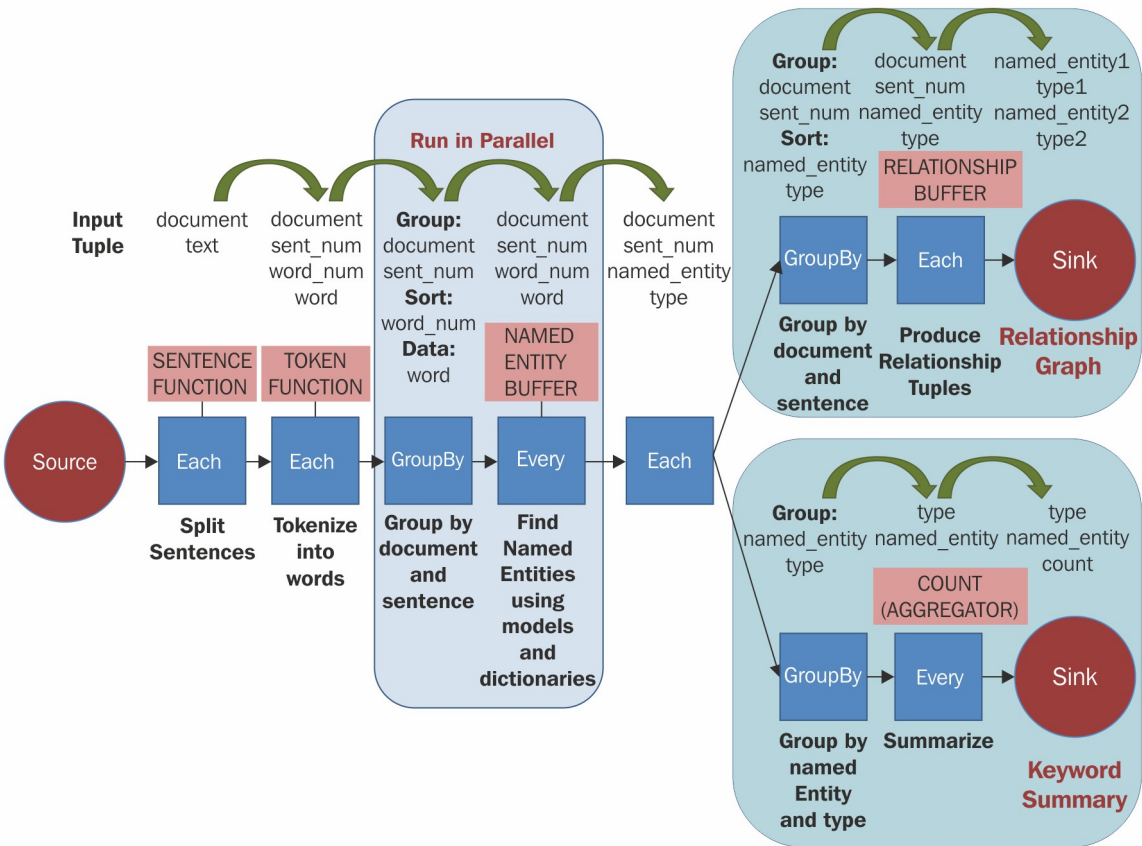
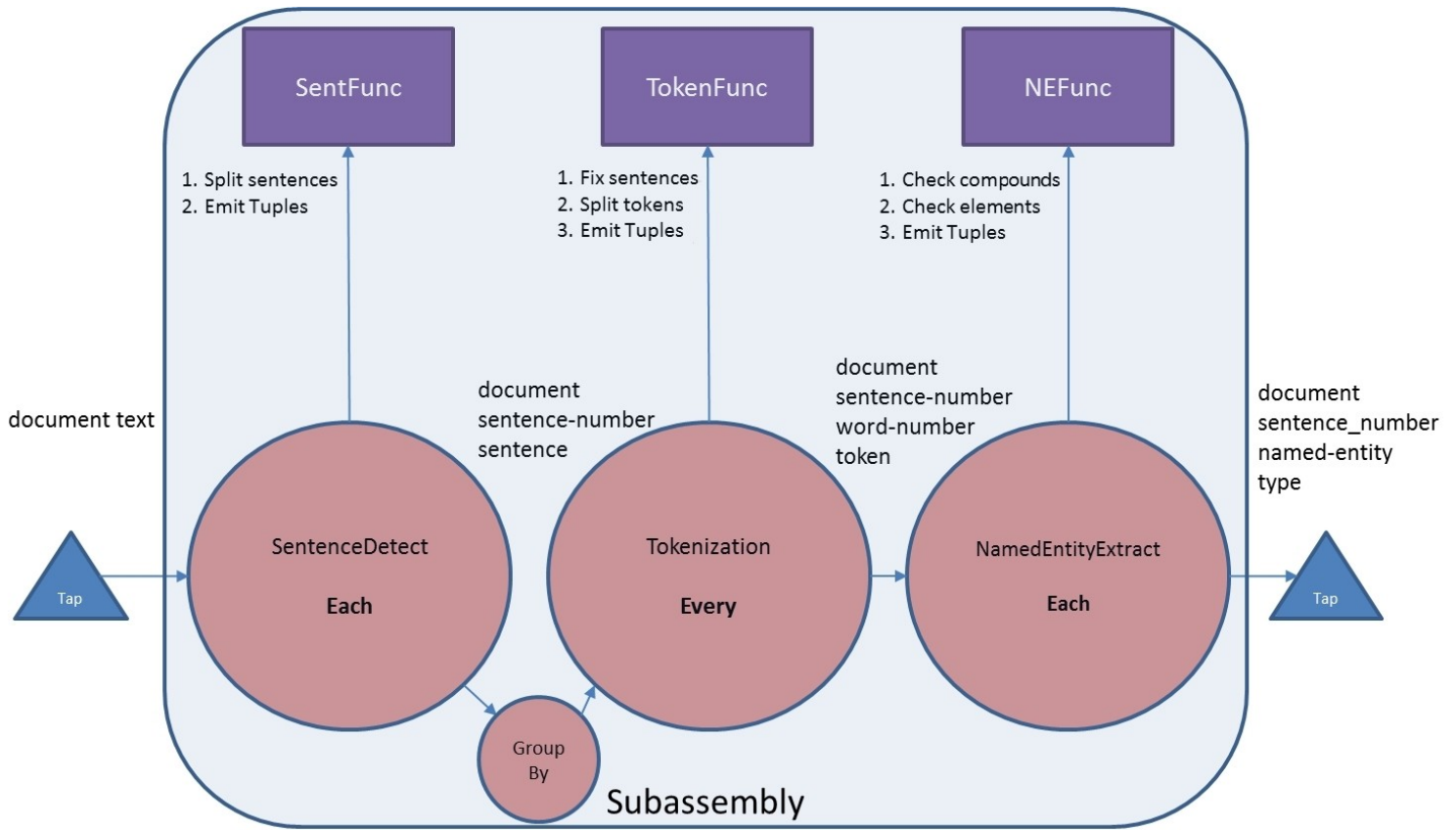


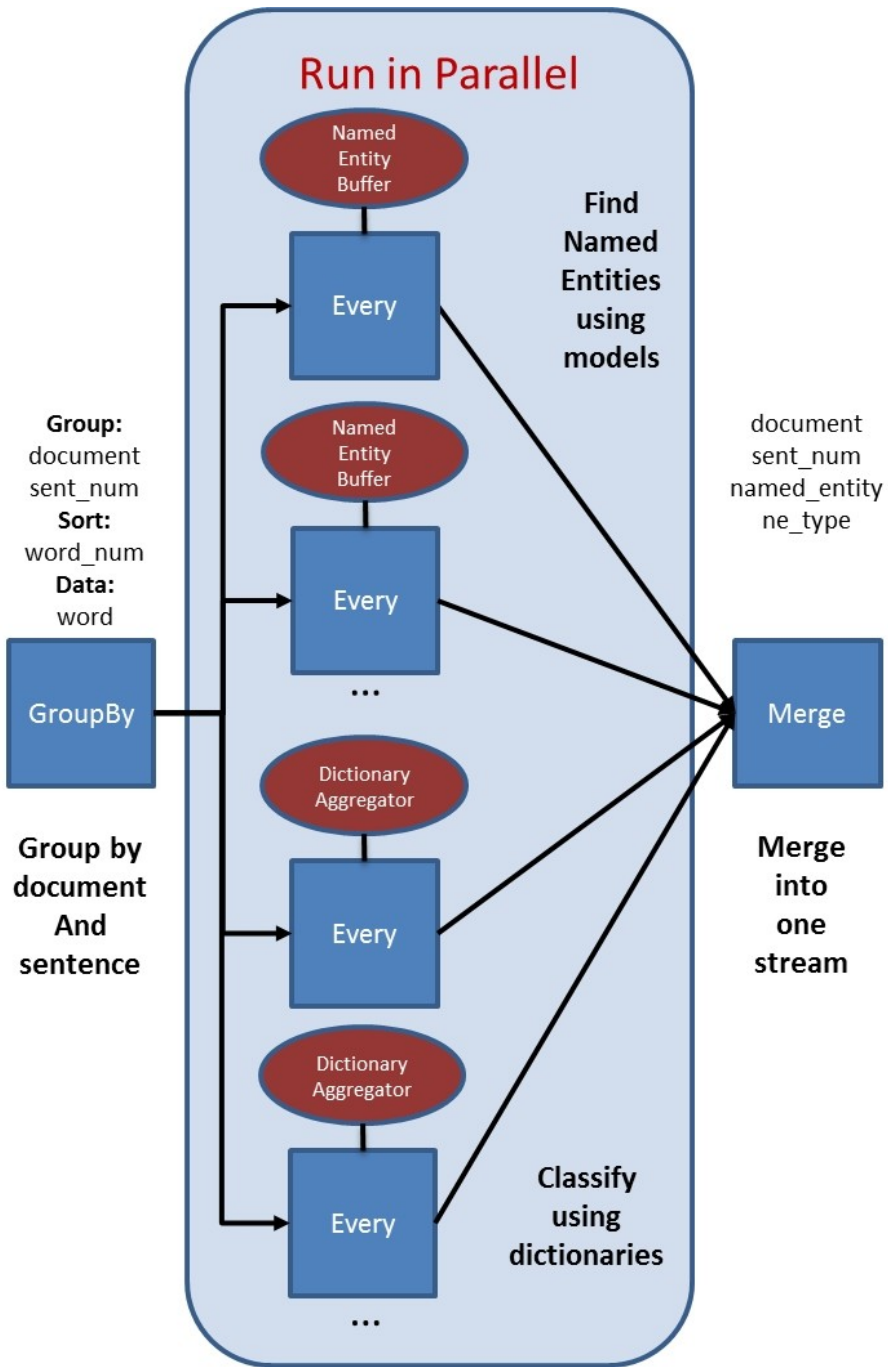
Chapter 7: Optimizing the Performance of a Cascading Application

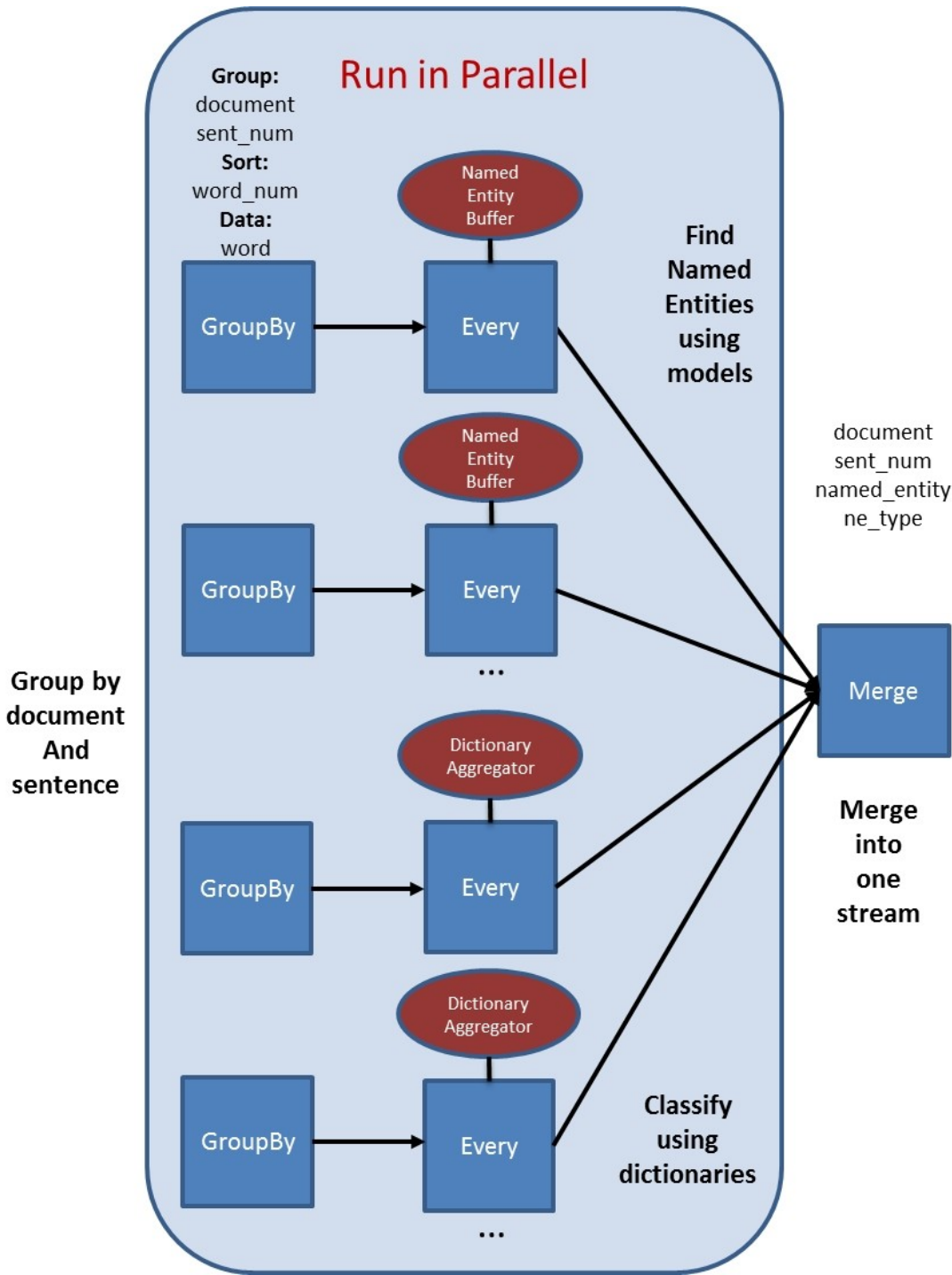


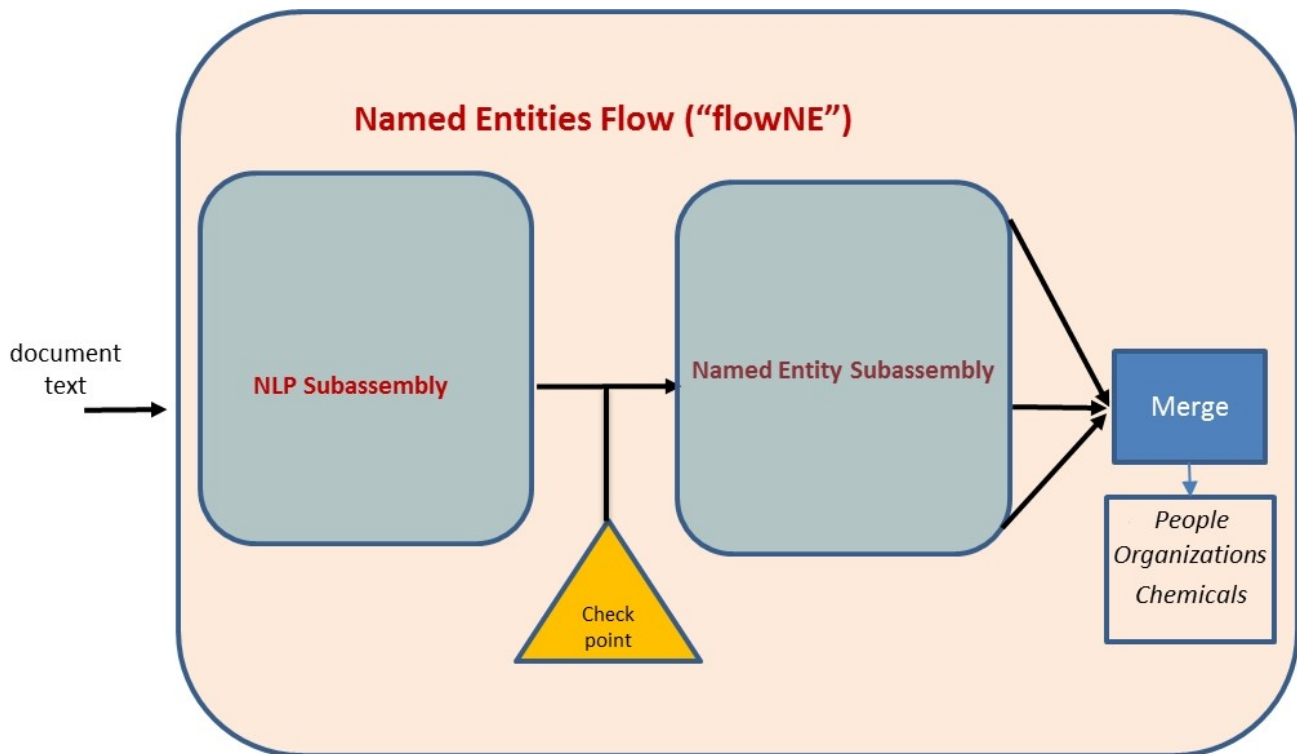
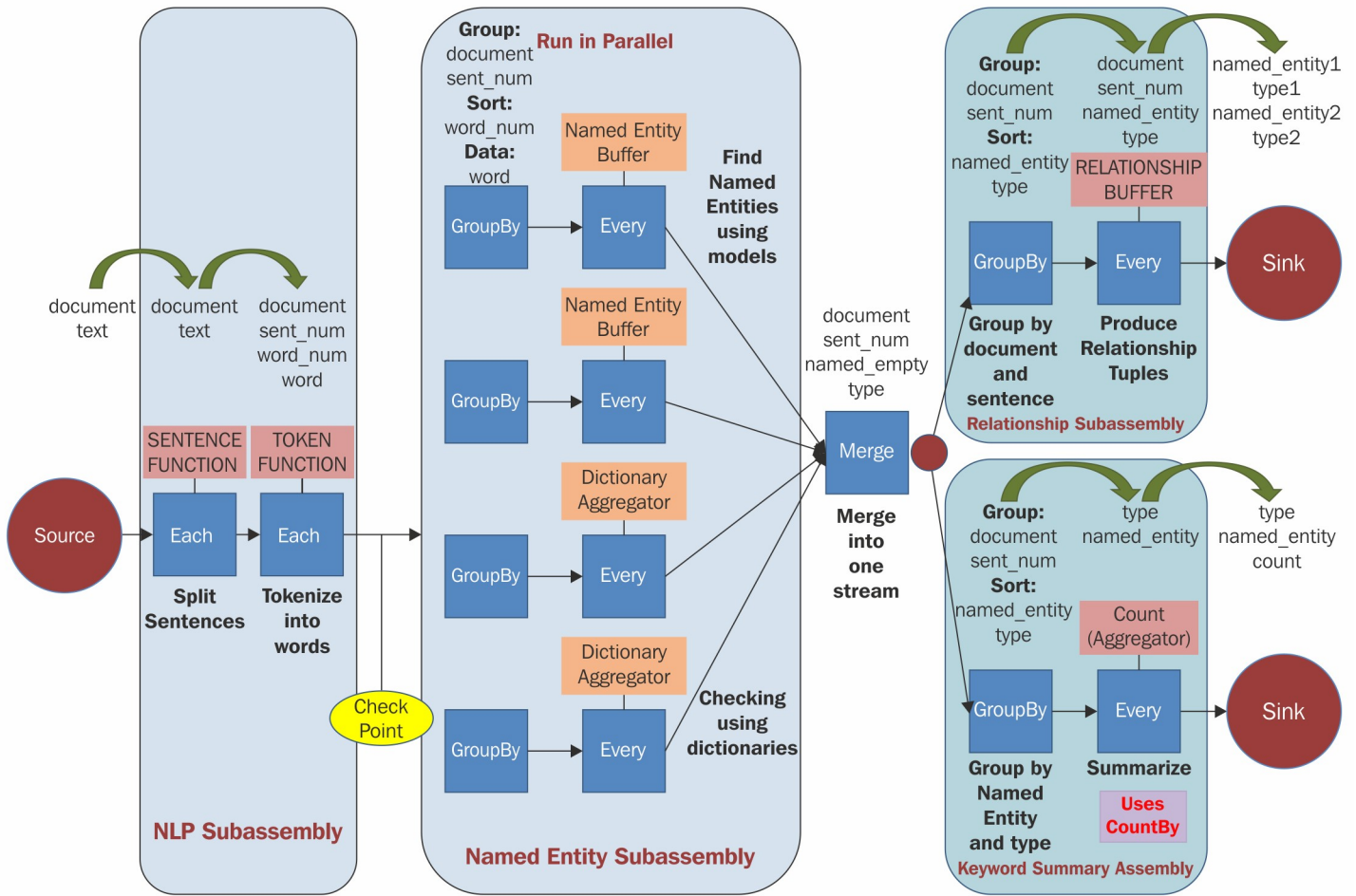
Compression format	File extension	Is It Splittable?
gzip	gz	No
bzip2	bz2	Yes
LZO	lzo	Yes (if indexed using Hadoop LZO indexing tool)
Snappy	snappy	No

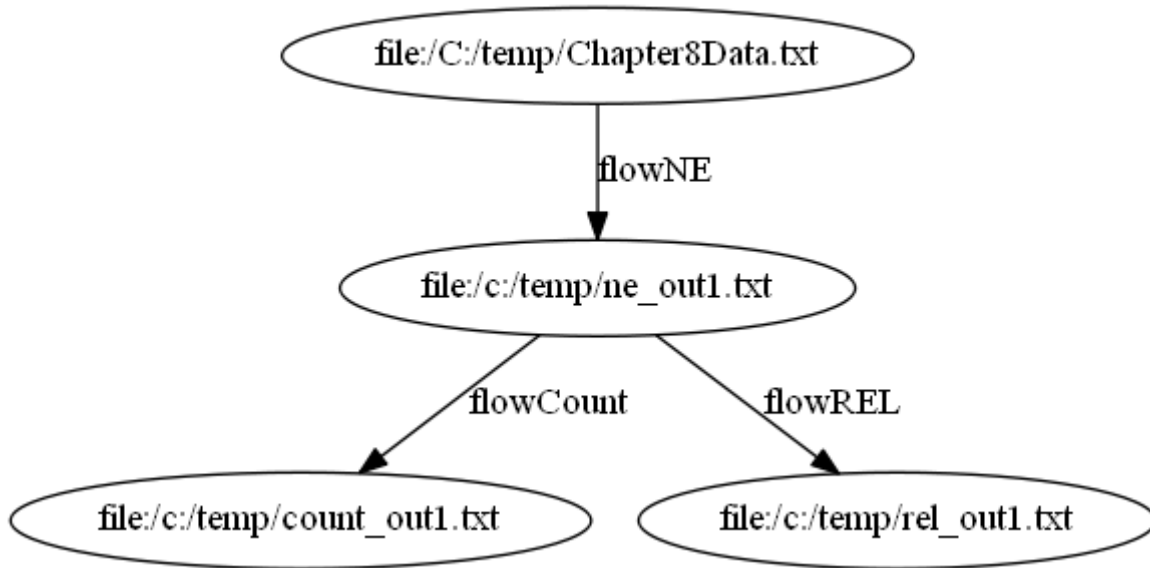
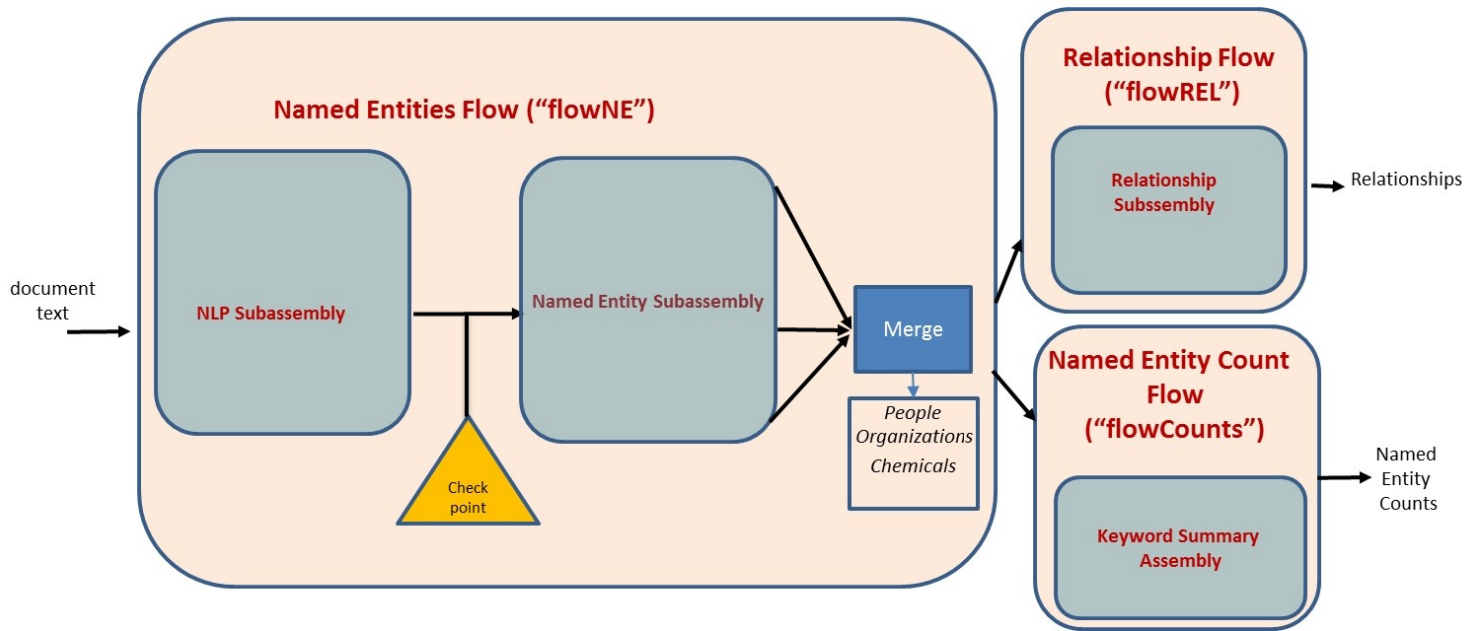
Chapter 8: Creating a Real-world Application in Cascading











Package Explorer JUnit
























UnitFuncTestJob

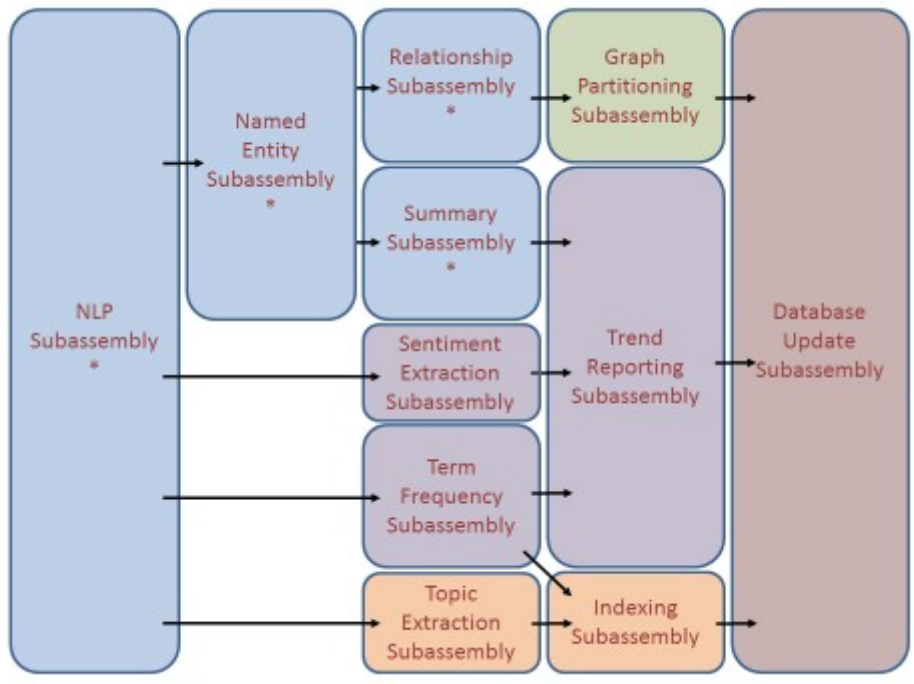
Runs: 1/1 Errors: 0 Failures: 0

com.ai.learning.jobs.UnitFuncTestJob [Runner: JUnit 4] (0.289 s)

- testSentenceFunction (0.289 s)

Failure Trace

-  cascading-core-2.6.1.jar
-  cascading-core-2.6.1-tests.jar
-  cascading-hadoop-2.6.1.jar
-  cascading-hadoop2-mr1-2.6.1.jar
-  cascading-local-2.6.1.jar
-  cascading-xml-2.6.1.jar
-  commons-lang-2.6.jar
-  guava-14.0.1.jar
-  hadoop-common.jar
-  hadoop-core.jar
-  hadoop-core-0.20.205.0.jar
-  hadoop-hdfs.jar
-  hadoop-tools.jar
-  hadoop-tools-0.20.205.0.jar
-  janino-2.6.1.jar
-  jgrapht-jdk1.6-0.8.1.jar
-  jwnl-1.3.3.jar
-  opennlp-maxent-3.0.3.jar
-  opennlp-tools-1.5.3.jar
-  riffle-0.1-dev.jar
-  slf4j-api-1.7.10.jar
-  slf4j-simple-1.7.10.jar
-  tagsoup-1.2.jar



Chapter 9: Planning for Future Growth

Programming Language	Project Name	Description of Project
Clojure	Cascalog	Clojure
Java	Cascading	Concurrent Cascading
JRuby	Cascading.JRuby	From Etsy, JRuby
PMML	Pattern	Concurrent PMML
	JPMML-Cascading	Open scoring PMML
Python	PyCascading	Twitter Python
Scala	Scalding	Twitter Scala
SQL	Lingual	Concurrent ANSI SQL shell and JDBC driver

Data Source	Project Name	Description of Project
Accumulo	Cascading.Accumulo	Accumulo data source
Cassandra	Cascading-Cassandra	Cassandra data source
Derby	Cascading-JDBC	Derby data source via JDBC
Elasticsearch	elasticsearch-hadoop	Elasticsearch data source
ElephantDB	ElephantDB	ElephantDB data source
H2	Cascading-JDBC	H2 data source via JDBC
HBase	Cascading.HBase	HBase data source
Hive	Cascading-Hive	Hive HQL
	Cascading.Hive	Hive data source
JDBC	Cascading-JDBC	Concurrent JDBC drivers
Oracle	Cascading-JDBC	Oracle database JDBC drivers
Memcached	Cascading.Memcached	Memcached data source
MongoDB	Cascading-Mongomigrate	MongoDB data source
MySQL	Cascading-JDBC	MySQL database JDBC drivers
Neo4j	Cascading.Neo4j	Neo4j data source
Parquet	Parquet-mr	Parquet data source
PostgreSQL	Cascading-JDBC	PostgreSQL database JDBC drivers
Redshift	Cascading-JDBC	Amazon Redshift database JDBC drivers
SimpleDB	Cascading.SimpleDB	Scale Unlimited SimpleDB data source
Solr	Cascading.Solr	Scale Unlimited Solr data source
Splunk	Tbana	Splunk data source
Teradata	Cascading-JDBC	Teradata database JDBC drivers

Serializer	Project Name	Description of Project
Avro	Cascading.Avro	Scale Unlimited data serialization for Apache Avro
JSON	Cascading.JSON	JavaScript Object Notation (JSON) utility classes
Kryo	Cascading.Kryo	Kryo serialization
Protocol Buffers	Cascading2-protobufs	Square Protocol Buffers
Thrift	Cascading-Thrift	Thrift Serializer

• Cascading-Hive Part3 TPC-DS [🔗](#)

App: **Cascading-Hive Part3 TPC-DS**

Owner: ryan | Jar Info: cascading-hive-1.0.0.jar | Platform: Hadoop | App Tags: | Frameworks: cascading-hive:1.0.0 | Run Time: 6m 37s | Progress: 12/12 steps | Tuples read: 659,682 | Tuples written: 86,019

