Hands on analysis of NGS data with Sparkhit

de.NBI summer school 2017
(30.06.2017)

Liren Huang
Jan Krüger
Supervisor: Alexander Sczyrba
Bielefeld University

The future is already here -
it is just not very evenly distributed

William Ford Gibson

Overview

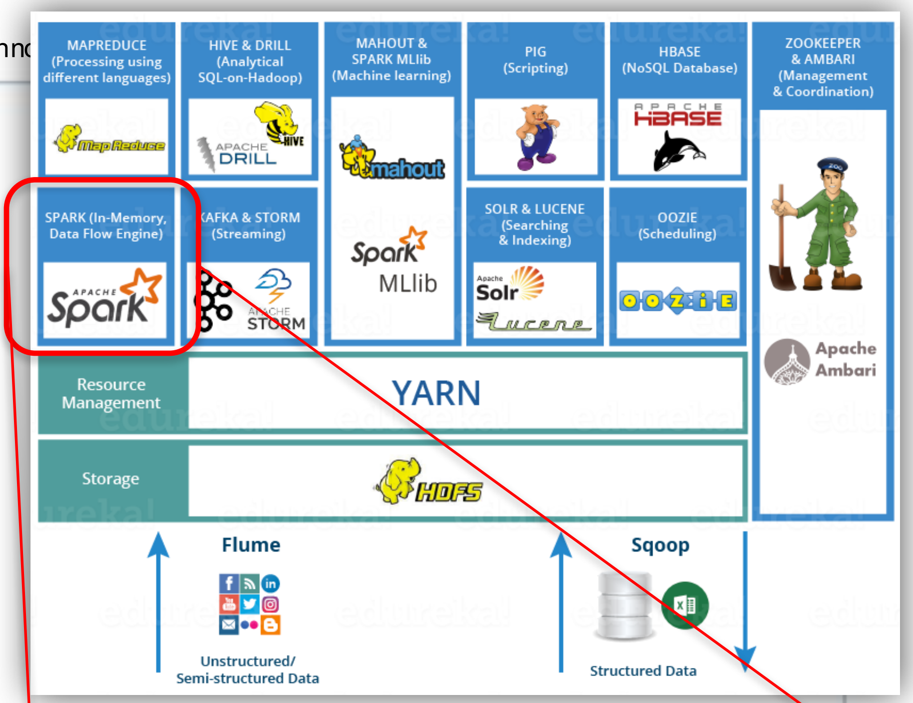
- **Introduction**
  - Apache Spark and RDD
  - Sparkhit, a toolkit for NGS data analysis

- **Hands on section**
  - Spark shell programming with RDD`s interface
  - Analyzing NGS data with Sparkhit

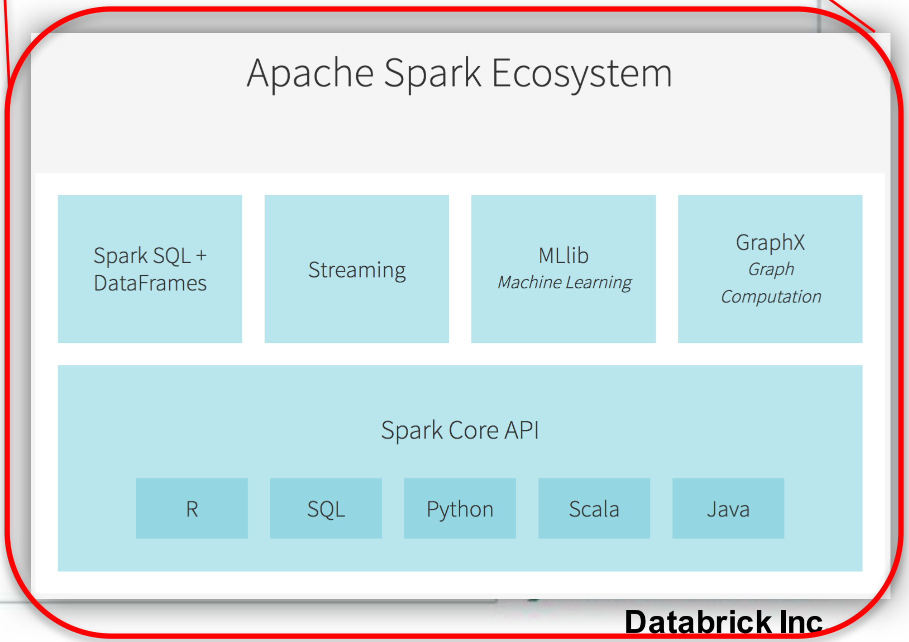
Apache Spark

is a fast and general engine for large-scale data processing

- An Extended Map-Reduce model
- A distributed programming engine that can interact with most tools in Hadoop eco-system
- Its core is a distributed data abstraction called RDD (resilient distributed dataset)


MAPREDUCE (Processing using different languages)
MapReduce
HIVE & DRILL (Analytical SQL-on-Hadoop)
APACHE DRILL
HIVE
MAHOUT & SPARK MLlib (Machine learning)
mahout
Spark MLlib
PIG (Scripting)
HBASE (NoSQL Database)
APACHE HBASE
ZOOKEEPER & AMBARI (Management & Coordination)
SPARK (In-Memory, Data Flow Engine)
APACHE Spark
KAFKA & STORM (Streaming)
APACHE STORM
SOLR & LUCENE (Searching & Indexing)
Apache Solr
Lucene
OOZIE (Scheduling)
OOZIE
Apache Ambari
Resource Management
YARN
Storage
HDFS
Flume
Sqoop
Unstructured/ Semi-structured Data
Structured Data

Edureka Inc.


Apache Spark Ecosystem

Spark SQL + DataFrames	Streaming	MLlib *Machine Learning*	GraphX *Graph Computation*

Spark Core API

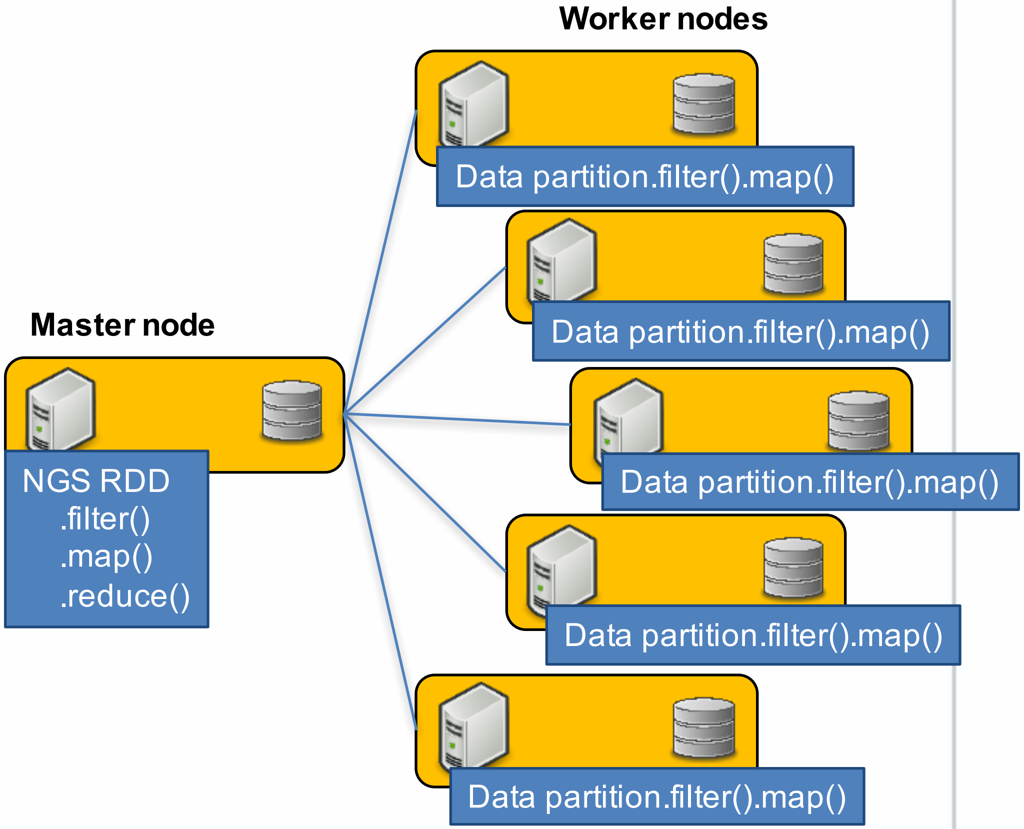
R SQL Python Scala Java

Databrick Inc.

RDD parallelization

- A RDD is an object.
- A RDD consists of several Data partitions across the cluster.
- An operation to RDD is parallelized to each partition

The benefit is you focus on your algorithm while Spark distributed the workload for you.


Worker nodes
Data partition.filter().map()
Data partition.filter().map()
Data partition.filter().map()
Data partition.filter().map()
Data partition.filter().map()
Master node
NGS RDD
.filter()
.map()
.reduce()

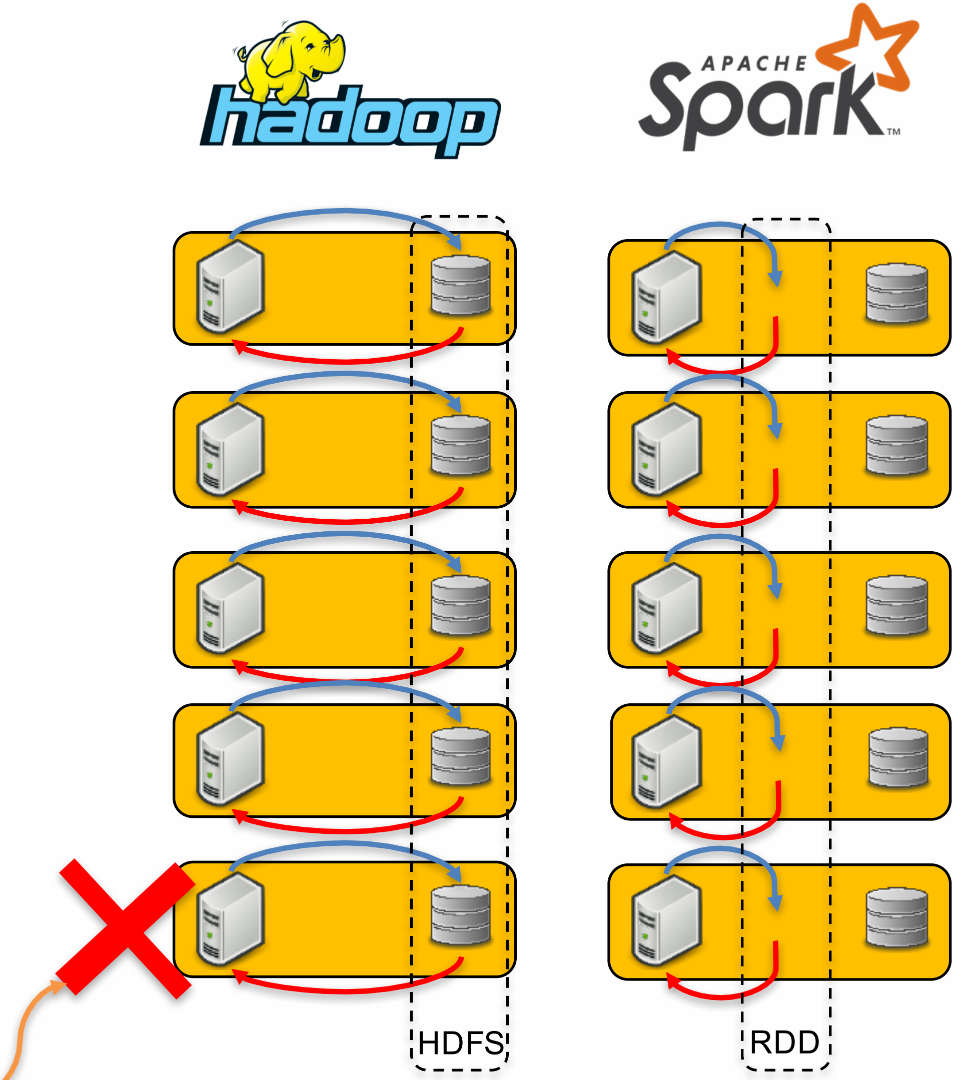
RDD cache

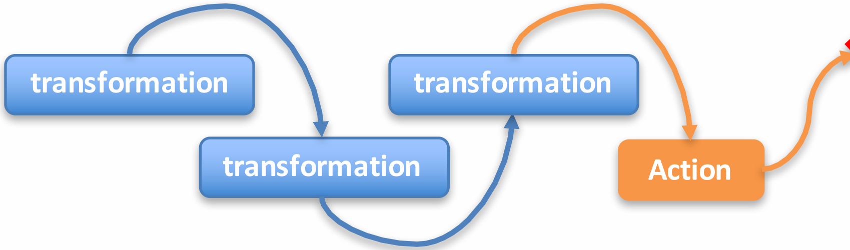
(Resilient distributed dataset)

- Distributed in memory computation for faster iterative algorithms

- Two types of operations
  - Transformation
  - Action

- Lazy feature (will see later), related to fault tolerance mechanism.





Sparkhit-mapper

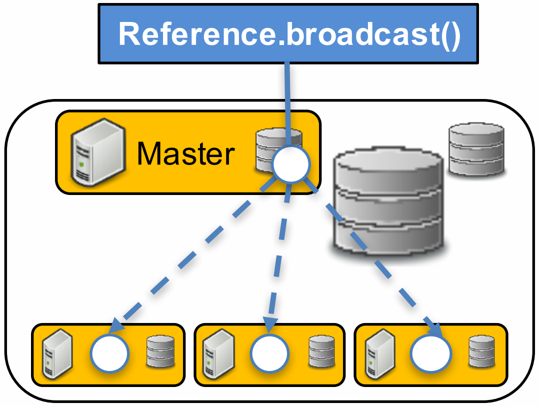
Sparkhit is a bioinformatics toolkit build on the Apache Spark platform

Here we describe a fragment recruitment application (short read mapping) call Sparkhit-mapper

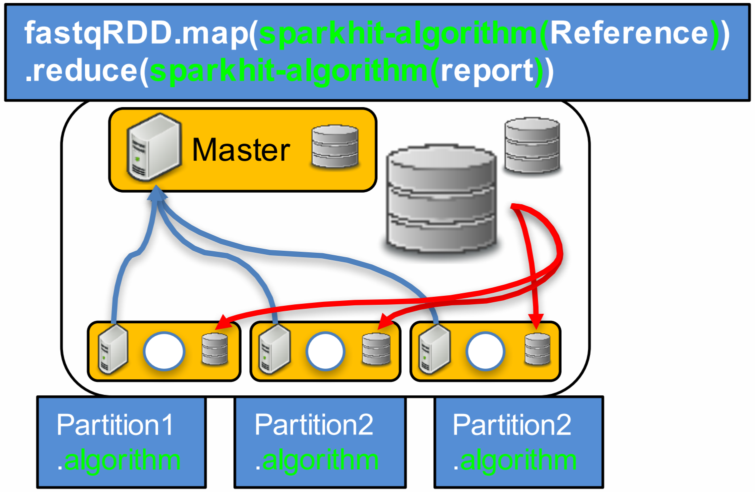
1, build reference index

2, broadcast to each worker nodes

3, each worker applies implemented alignment-algorithm for recruiting the fragments.



Reference.broadcast()



fastqRDD.map(sparkhit-algorithm(Reference))
.reduce(sparkhit-algorithm(report))

https://rhinempi.github.io/sparkhit/

Acknowledgement

- Dr. Alexander Sczyrba, de.NBI Bielefeld
- Jan Krüger, de.NBI Bielefeld
- Dr. Burkhard Linke, de.NBI Giessen

Online tutorial:

https://rhinempi.github.io/sparkhit/usecase.html