

Uvod

Apache Spark

Apache Spark

- Apache Spark je “*fast, general-purpose, distributed big data processing platform*”.
- Efikasna upotreba memorije
 - Do 100 puta brži od sličnih platformi
- API zasnovan na skupu collection-based procedura koji sakriva da se radi sa klasterom mašina

Osobine

- Distribuirana platforma, ali sa mogućnošću da se velike količine podataka u memoriji
- Spark kolekcija sakriva činjenicu da se referenciraju podaci sa različitih čvorova
- Podržava
 - Batch programiranje, real-time procesiranje, SQL-like upite, algoritme na grafovima, algoritme mašinskog učenja
 - Python, Java, Scala, R
- Nije pogodan za OLTP obradu ili kada dataset može da obradi jedna mašina

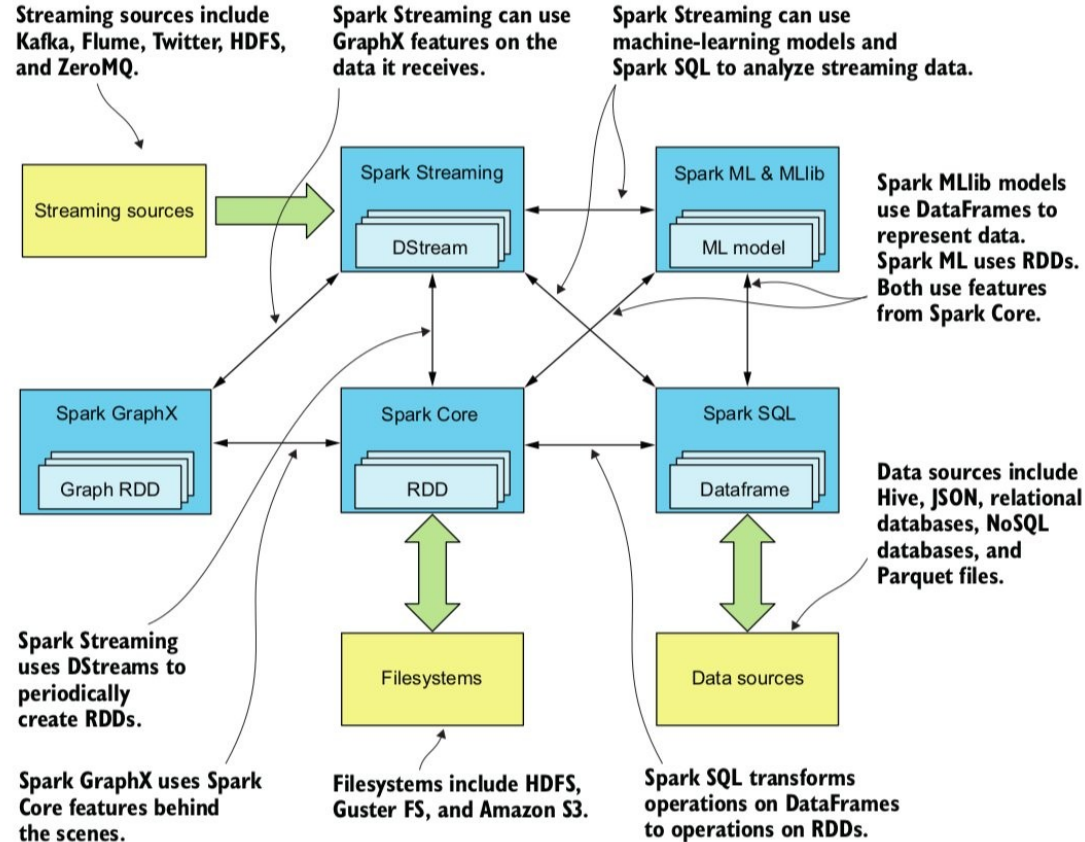
Hadoop

- Osobine
 - Paralelizam
 - Distribuiranost
 - Fault tolerance
- Hadoop =
 - HDFS (Hadoop Distributed File System)
 - MapReduce data-processing engine, pri čemu rezultat jednog koraka mora da bude sačuvan u HDFS da bi ga upotrijebio sljedeći korak u postupku obrade

Spark + Hadoop

- Jedinstvena platforma uz bogat API
- Keširanje podataka koji nastaju kao rezultat međukoraka
 - In-memory execution model
- Sortiranje 100TB podataka za manje od 1.5 sekundi (sortbenchmark.org)

Spark komponente



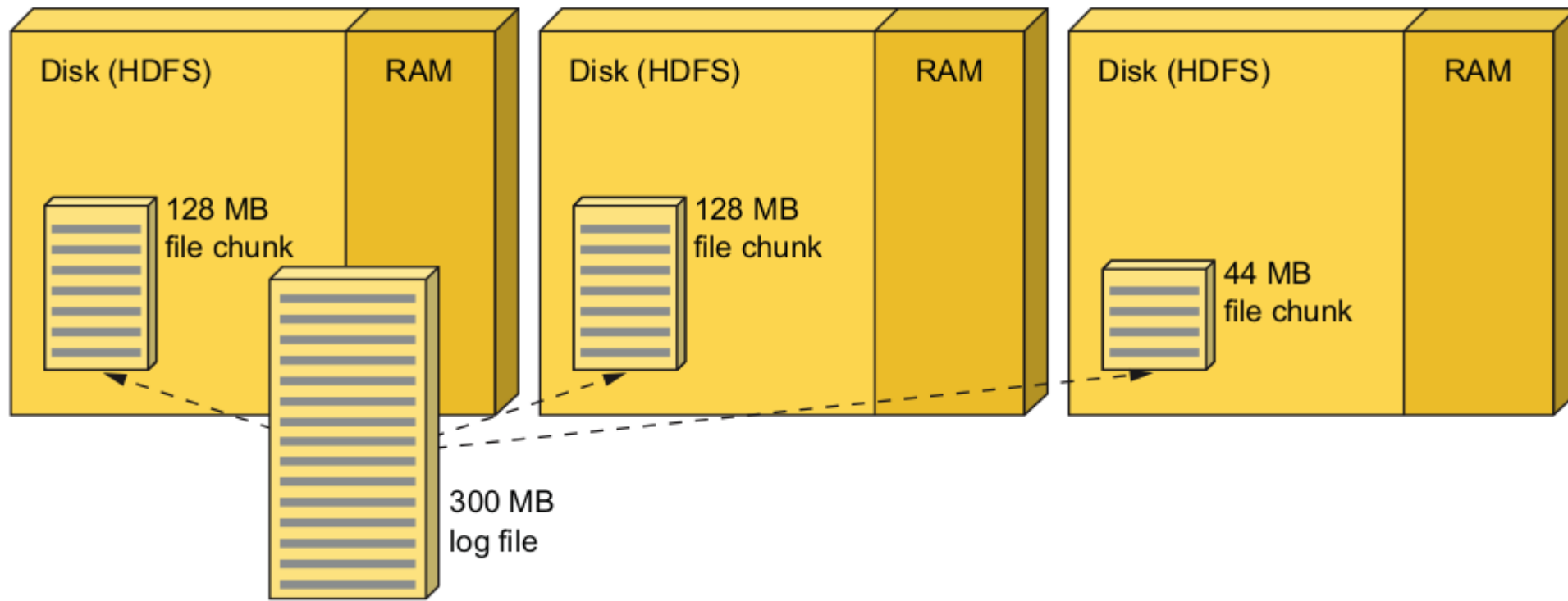
Spark komponente 2

- Spark Core, Resilient Distributed Dataset – RDD je apstrakcija distribuirane kolekcije
 - Immutable, resilient, distributed
 - Transformacije i akcije
- Spark SQL sa Dataframe-ovima omogućava manipulisanje velikim skupovima distribuiranih strukturiranih podataka. Operacije nad DataFrame-ovima mapiraju se na RDD operacije

Spark komponente 3

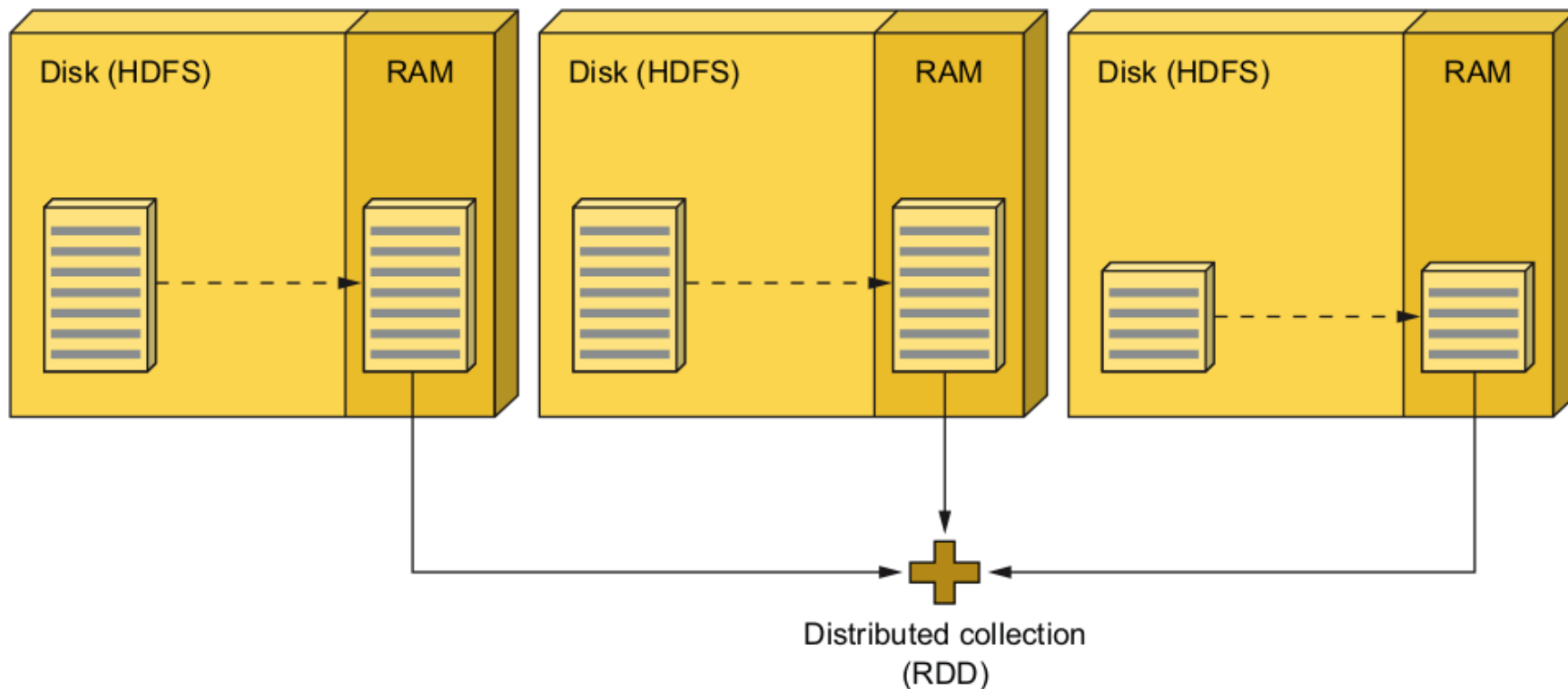
- Spark streaming – procesiranje real-time strimova podataka. Izvori mogu da budu razni, HDFS, ZeroMQ, Twitter itd. Dstream je RDD koji sadrži podatke iz posljednjeg “prozora”
- Spark Mllib je biblioteka za algoritme mašinskog učenja
- Spark GraphX sadrži metode za kreiranje i rad sa graph RDD-ovima.

Spark program flow



Spark program flow 2

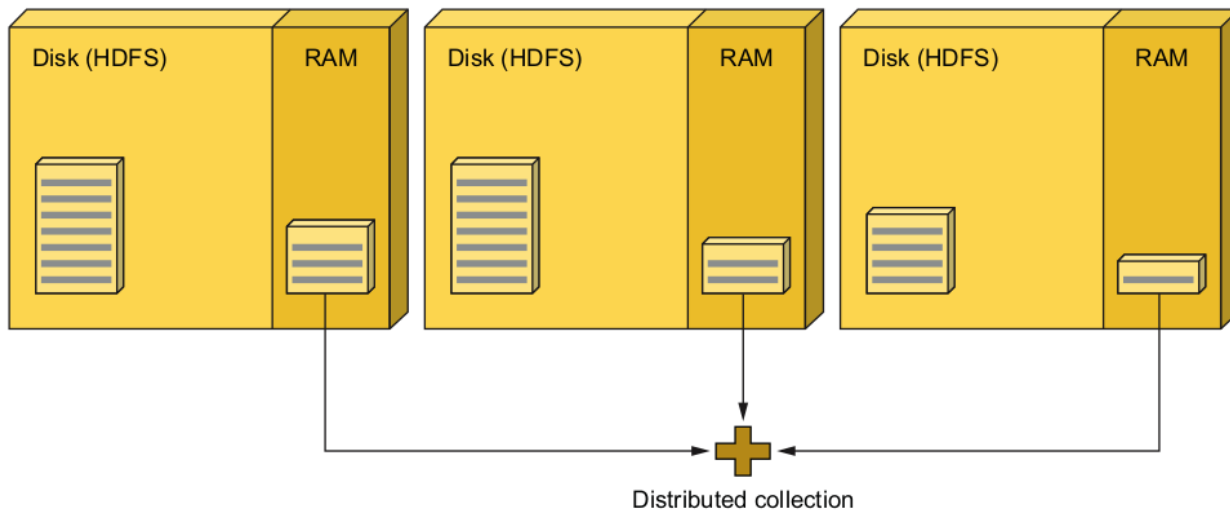
```
val lines = sc.textFile("hdfs://path/to/the/file")
```



Spark program flow 3

- Koliko se grešaka tipa *OutOfMemory* generisalo tokom posljednje dvije sedmice?

```
val oomLines = lines.filter(l => l.contains("OutOfMemoryError")).cache()
```



```
val result = oomLines.count()
```