Iterative Computing on Massive Data Sets

William Rory Kronmiller

What Is It?

Spark The Distributed System

- UC Berkeley Research Project
- Designed to run on a cluster of unreliable machines
- Resilient Distributed Datasets (RDDs)
  - Collections
  - Partitioned and Parallel
  - Immutable
  - Recoverable
  - **LAZY**
    - **Materialized by output operators***
- In-Memory

Why Do I Care?

Some Companies Using Spark

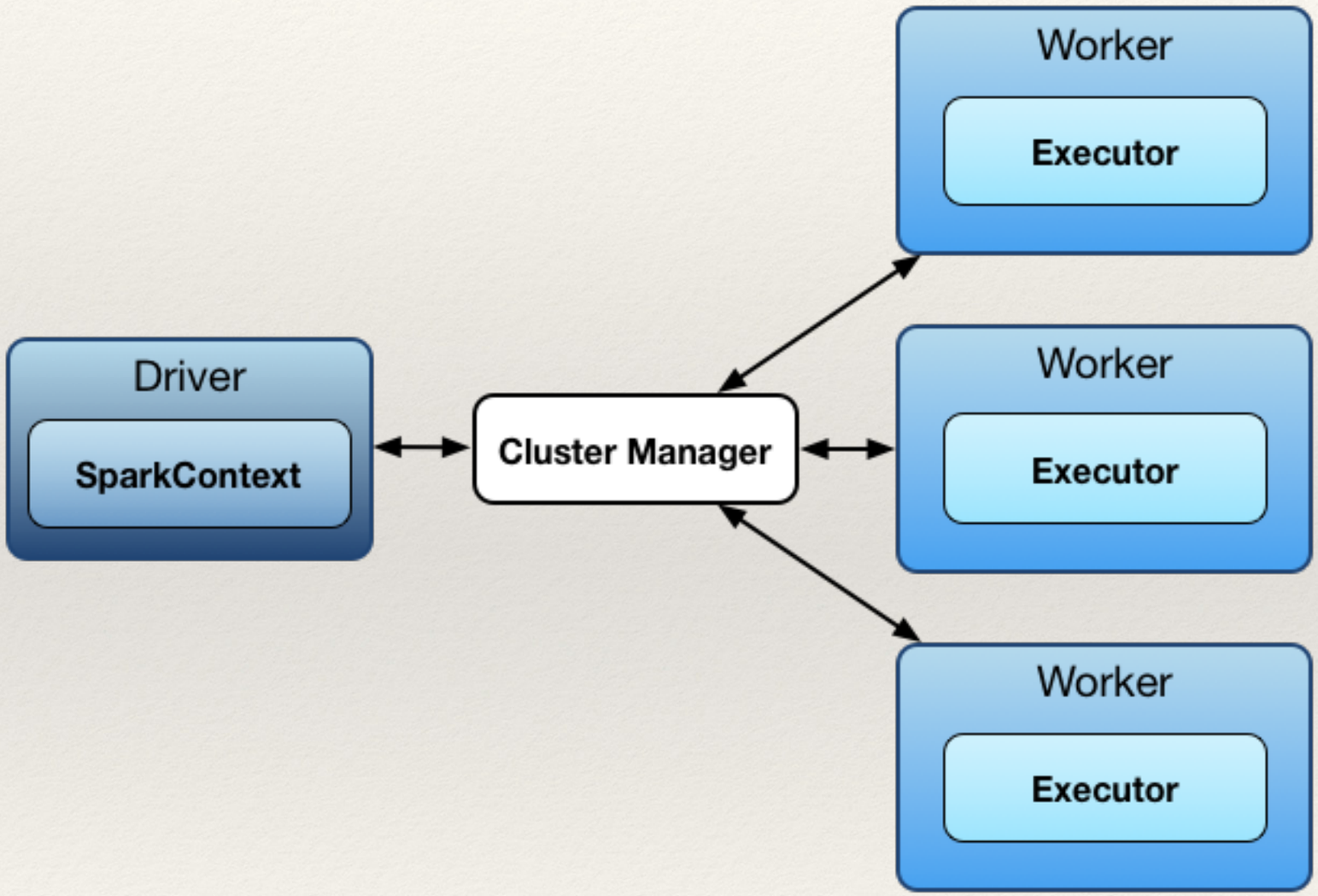
- Alibaba
- Amazon
- Baidu
- Berkeley AMPLab
- CERN
- Concur
- Databricks
- eBay
- Facebook
- NASA JPL
- Opentable
- Tencent
- TripAdvisor

Some Things You Can Do With Spark

- Real-Time Advertisement Selection
- Genome Sequencing
- Network IDS
- Fraud Detection

https://www.mapr.com/blog/game-changing-real-time-use-cases-apache-spark-on-hadoop

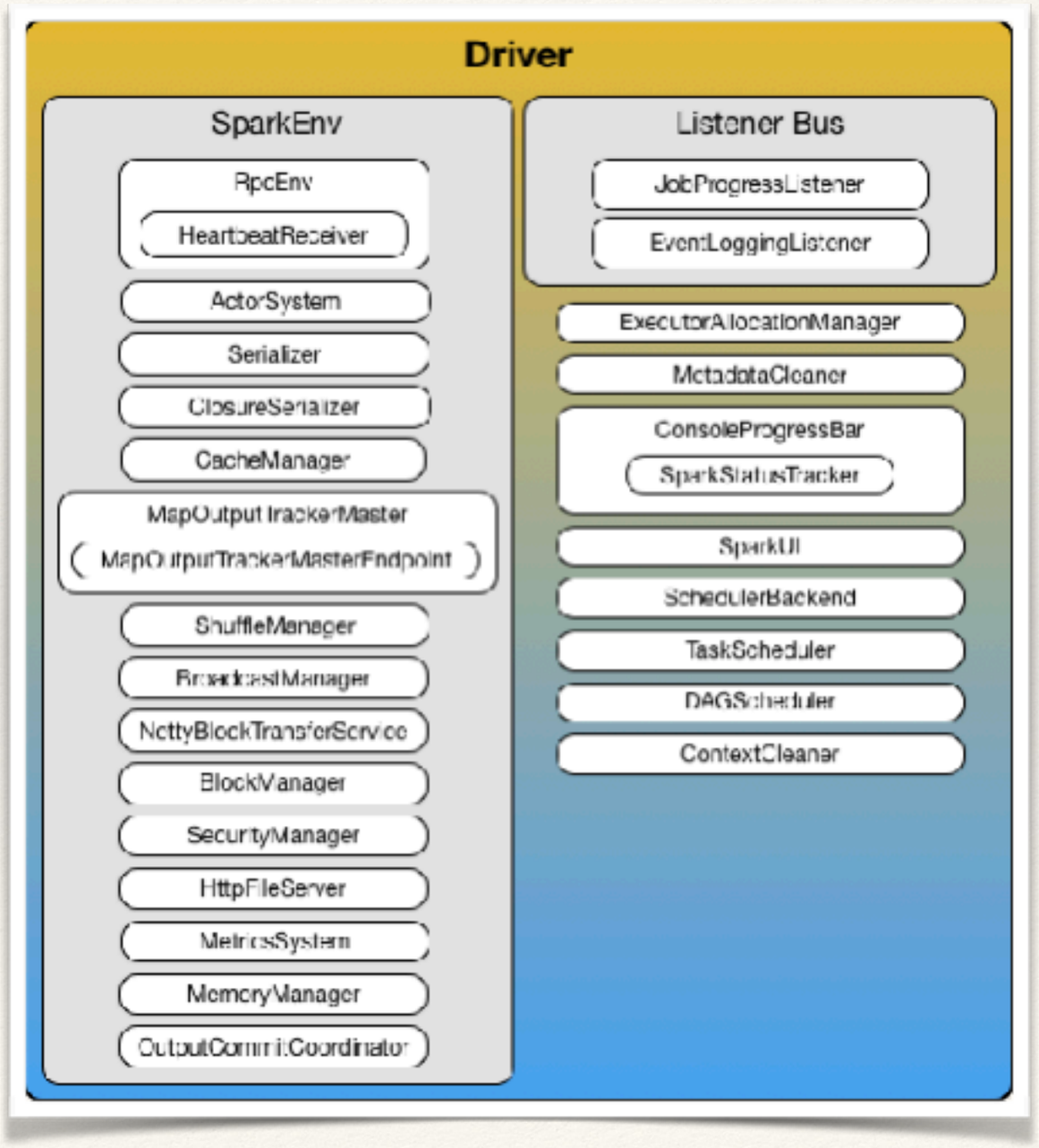
Spark High Level Design



https://jaceklaskowski.gitbooks.io/mastering-apache-spark/content/spark-architecture.html

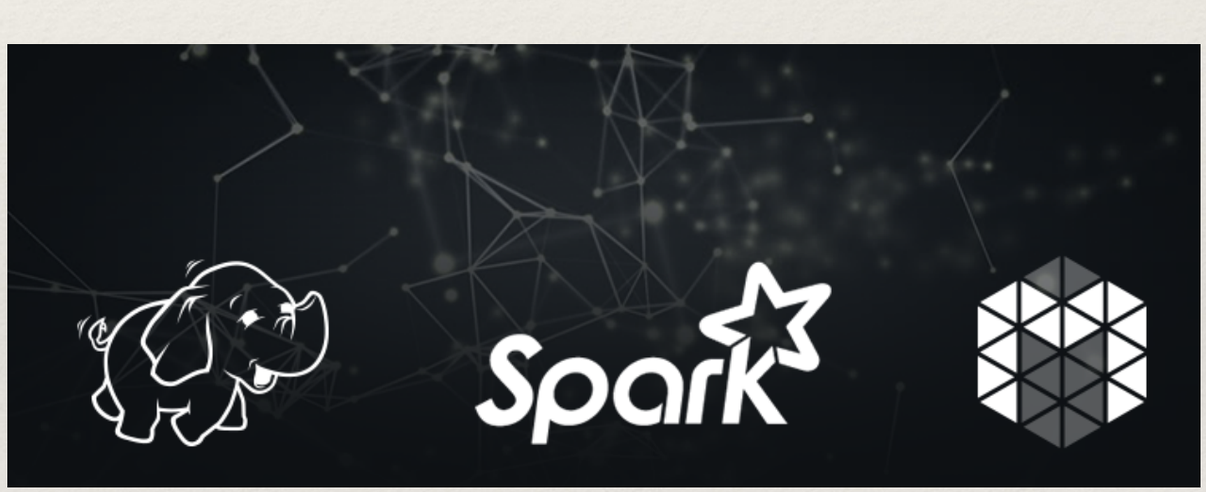
Spark Driver

- The Core of a Spark Job:
  - Talks to cluster
  - Creates and schedules tasks
  - Tracks computation progress
  - Hosts WebUI



Cluster Manager

- Spark runs on multiple clusters:
  - Mesos
  - YARN
  - Spark Standalone



http://www.agildata.com/wp-content/uploads/2016/03/Blog-InlineIMAGES-ClusterManagement.png

Spark Executor

- Hosts RDD Partitions
- Performs Computations
- Ideally Executor Runs on HDFS Host for Partition

Basic Programming Model

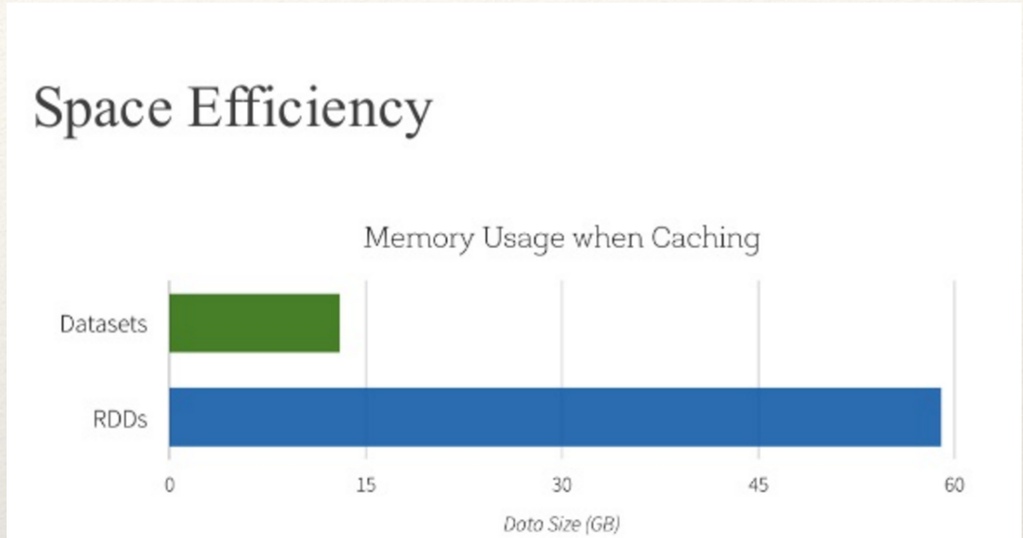
- Start with a collection
  - HDFS File(s)
  - Scala Collection
  - Database Table/Collection
- Do Transformations
  - map()
  - join()
  - filter()
- Output results…

Output Operators

- Trigger Materialization
- Sometimes called Output Operations (DStreams)
- Also called Actions (RDDs)
- Examples
  - reduce()
  - collect()
  - count()
  - foreach()
  - take()
  - saveAsTextFile()

Spark SQL

- RDDs as SQL Tables
  - DataFrame
    - Similar to R DataFrames ¯_(ツ)_/¯
    - Untyped :(
  - DataSet
    - Typed :)
    - Has Become Superset for DataFrame
- Compatible with Hive and HiveQL
- Compatible with JDBC and ODBC thru Thrift Server



	SQL	DataFrames	Datasets
Syntax Errors	Runtime	Compile Time	Compile Time
Analysis Errors	Runtime	Runtime	Compile Time

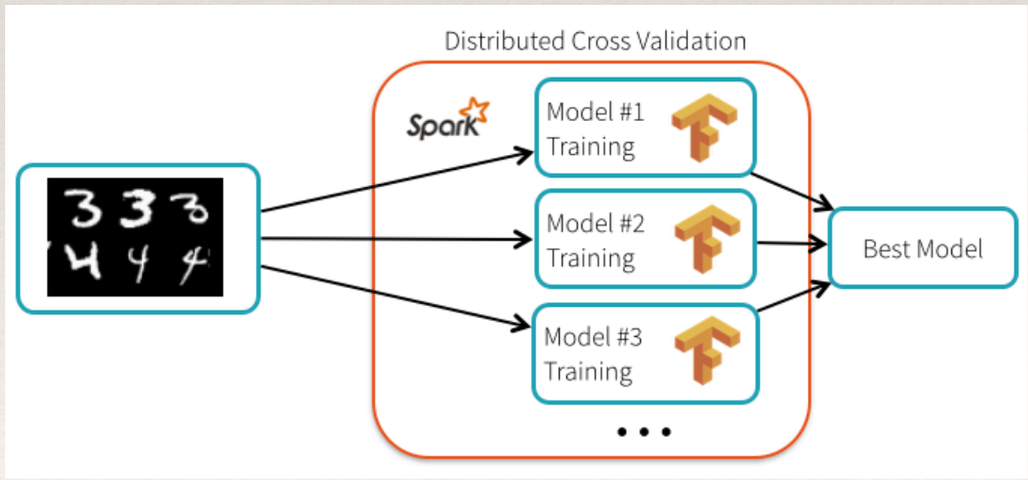
https://databricks.com/blog/2016/07/14/a-tale-of-three-apache-spark-apis-rdds-dataframes-and-datasets.html

Databricks Walk-Through

MLlib

- Machine Learning Library
- Huge Just Google It
- https://goo.gl/5m0ijJ
- https://goo.gl/uM5CLv
- Could Be Its Own UPE IO

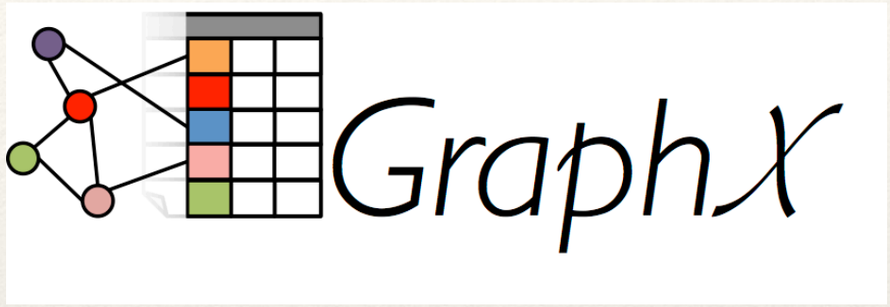
https://cdn.meme.am/instances/500x/67240542.jpg



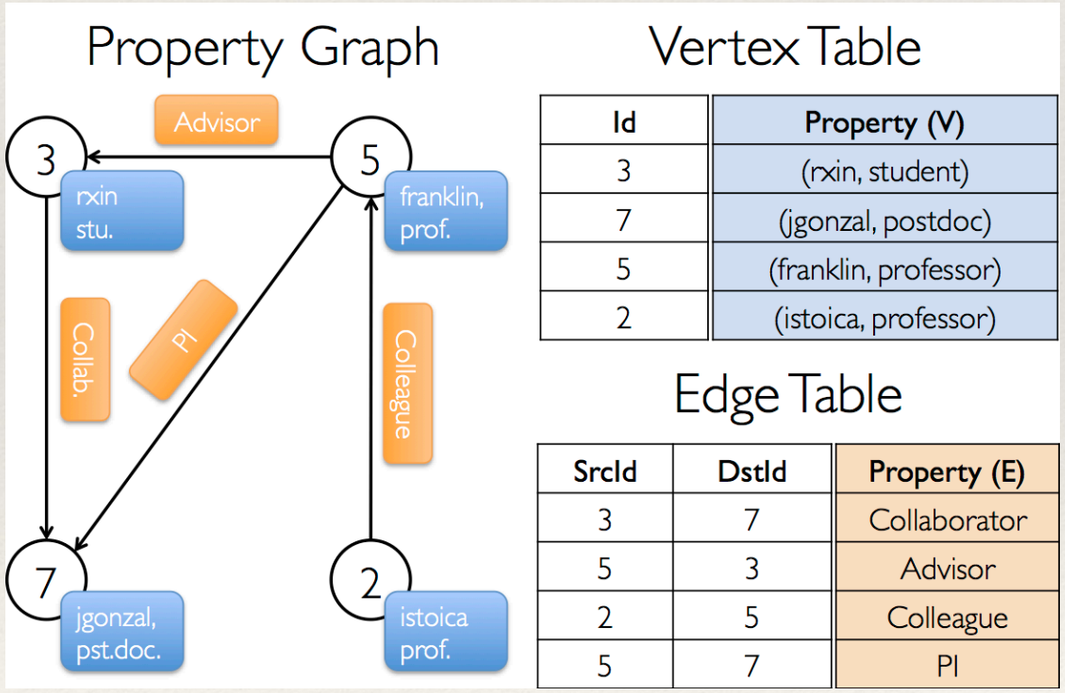
https://databricks.com/wp-content/uploads/2016/01/image04.png

GraphX

- Graph Processing on Spark
- Edge Computation
- Vertex Centric Computation
  - Pregel API
- https://goo.gl/YM0ruw



http://spark.apache.org/docs/latest/img/graphx_logo.png



http://spark.apache.org/docs/latest/img/property_graph.png

Postgres vs Spark

- SQL Databases Are Good For
  - Transactions
  - Low Latency
  - Small-Medium Size Data Sets
  - One Big Server
  - (Probably) Efficiency: Decades of Query Optimizers
- Spark is Good For
  - ETL
  - Data Mining
  - Machine Learning
  - Cloud Environments

- PostgreSQL
  - Unlimited Database Size
  - 32TB Table Size
  - 1.6TB Row Size
- Spark
  - 2014 Daytona GraySort Contest Winner
    - 100TB in 23 Minutes
    - 1000TB in 243 Minutes

Hadoop MapReduce vs Spark

- Both perform batch computation
- Both support Map/Reduce paradigm
- Both distributed/fault-tolerant
- Both use HDFS
- Spark RDDs Enable Iterative Computation

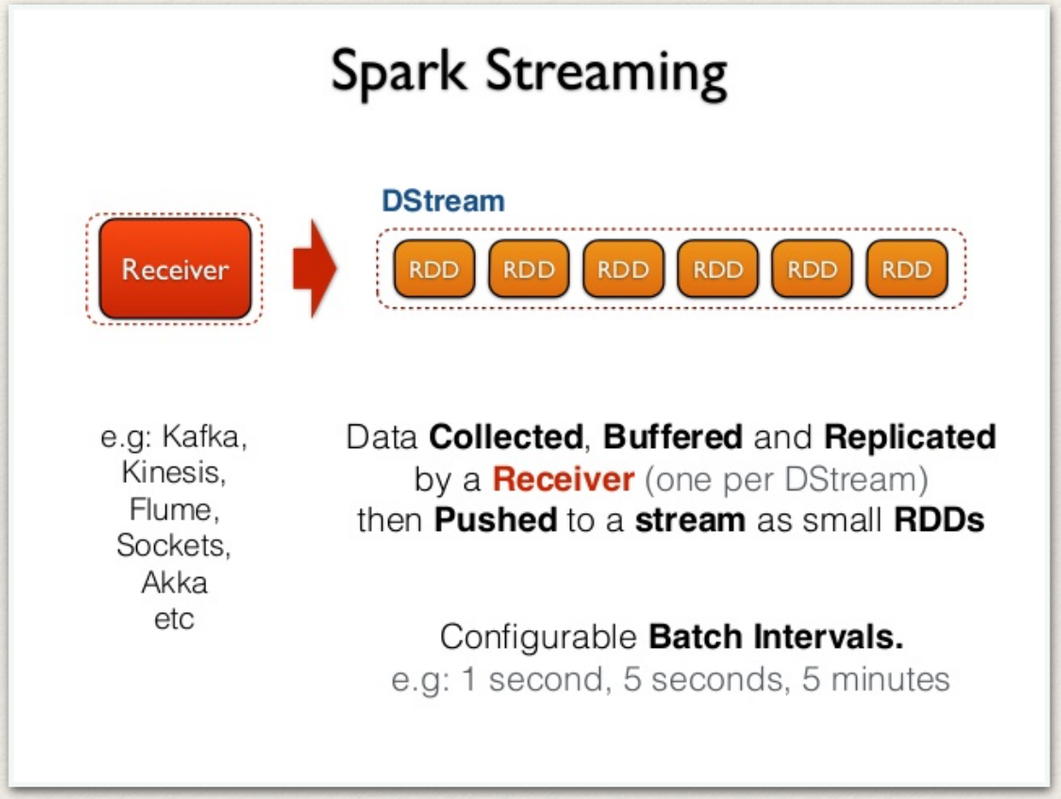
Spark Streaming

Micro Batch Model

- Spark Uses Batch Computation
  - Take a bucket of data
  - Transform the bucket of data
  - …
  - Profit
- What about unbounded streams of data?
  - Micro-Batches
    - Take a bucket of data every N seconds…

DStreams

- Micro-batch programming abstraction
- Unbounded collection of RDDs
  - Still parallel, fault-tolerant, etc…
  - Still lazy



http://image.slidesharecdn.com/apachesparkhs-141212054258-conversion-gate01/95/apache-spark-streaming-46-638.jpg?cb=1418381382

Streaming DEMO

Try This At Home

https://databricks.com/try-databricks

- spark.apache.org
- cwiki.apache.org/confluence/display/SPARK/Powered+By+Spark
- databricks.com
  - https://databricks.com/blog/2016/07/14/a-tale-of-three-apache-spark-apis-rdds-dataframes-and-datasets.html
  - https://docs.cloud.databricks.com/docs/latest/databricks_guide/index.html#00%20Welcome%20to%20Databricks.html
  - https://databricks.com/blog/2014/10/10/spark-petabyte-sort.html
  - https://databricks.com/blog/2016/05/24/genome-sequencing-in-a-nutshell.html
  - http://cdn2.hubspot.net/hubfs/438089/notebooks/Samples/Miscellaneous/Genome_Variant_Analysis_using_k-means_ADAM_and_Apache_Spark.html
  - https://databricks.com/blog/2014/08/14/mining-graph-data-with-spark-at-alibaba-taobao.html
- sparkhub.databricks.com
- deepspace.jpl.nasa.gov
- mapr.com/blog/game-changing-real-time-use-cases-apache-spark-on-hadoop
- databricks.com/blog/2016/08/31/apache-spark-scale-a-60-tb-production-use-case.html
- jaceklaskowski.gitbooks.io/mastering-apache-spark/content/spark-architecture.html
- people.csail.mit.edu/matei/papers/2010/hotcloud_spark.pdf
- people.csail.mit.edu/matei/papers/2012/nsdi_spark.pdf
- mapr.com/blog/how-get-started-using-apache-spark-graphx-scala
- blog.cloudera.com/blog/2014/03/why-apache-spark-is-a-crossover-hit-for-data-scientists/
- datamation.com/data-center/hadoop-vs.-spark-the-new-age-of-big-data.html
- aws.amazon.com/emr/details/spark/
- amplab.cs.berkeley.edu/publication
- www.oreilly.com/learning/apache-spark-for-atom-smashing-experiments
- www.ebaytechblog.com/2014/05/28/using-spark-to-ignite-data-analytics/
- spark-summit.org/2015/events/using-data-science-to-transform-opentable-into-your-local-dining-expert/
- code.facebook.com/posts/1671373793181703/apache-spark-scale-a-60-tb-production-use-case/