# Distributed Computing

DIANE WOODBRIDGE, PH.D

# GHC 2017

October 4 ~ October 6, 2017.

Orlando, Florida.

CALENDAR

## 2017 Grace Hopper Celebration of Women in Computing

SHARE THIS

The 2017 Grace Hopper Celebration of Women in Computing will take place on Wednesday, October 4 through Friday, October 6. GHC 17 will be held in Orlando, Florida. Hope to see you there!

**LOOKING FOR A JOB?**

Submit your resume

**HAVE A QUESTION?**

Contact us

**ABOUT THE GRACE HOPPER CELEBRATION**

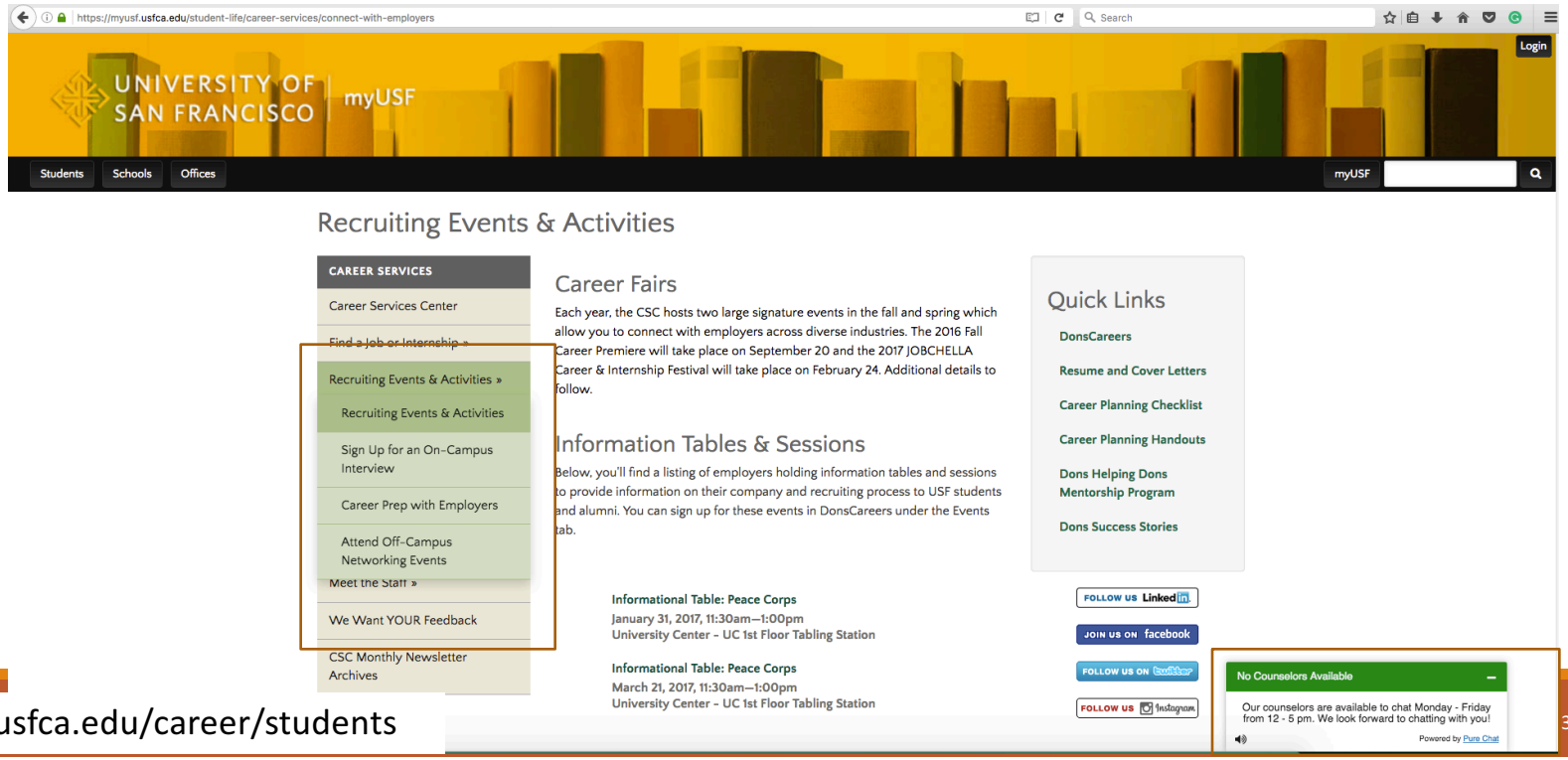
The Grace Hopper Celebration of Women in Computing (GHC) is the world's largest gathering of women technologists. It is produced by the Anita Borg Institute and presented in partnership with ACM.

Association for Computing Machinery

The Grace Hopper Celebration of Women in Computing is presented by the Anita Borg Institute for Women in Technology and the Association for Computing Machinery.

# Other career fairs?

USF, San Francisco, many more!




https://www.usfca.edu/career/students

# Distributed Computing

# Course Objectives

Understand needs and concepts of distributed computing.

Understand the Spark and its stack.

Being competent to work with Spark on a distributed computing environment.

- Programming with RDDs.
- Work with key/value pairs.
- Work with DataFrame.
- Work with SparkSQL, MLlib, ML, Spark Streaming, (and GarphX).
- Work on Amazon AWS.

# Spark Interview Questions

- ~~What is Apache Spark?~~
- ~~Explain the key features of Spark.~~
- ~~What is RDD?~~
- ~~How to create RDD.~~
- ~~What is ”partitions”?~~
- ~~Types or RDD operations?~~
- ~~What is “transformation”?~~
- ~~What is “action”?~~
- ~~Functions of “spark core”?~~
- ~~What is “spark context”?~~
- ~~What is an “RDD lineage”?~~
- ~~Which file systems does Spark support?~~
- ~~List the various types of “Cluster Managers” in Spark.~~
- ~~What is “YARN”?~~
- ~~What is “Mesos”?~~
- ~~What is a “worker node”?~~
- ~~What is an “accumulator”?~~
- ~~What is “Spark SQL” (Shark)?~~
- What is “SparkStreaming”?
- What is “GraphX”?
- ~~What is “MLlib”?~~



https://www.edureka.co/blog/interview-questions/top-apache-spark-interview-questions-2016/

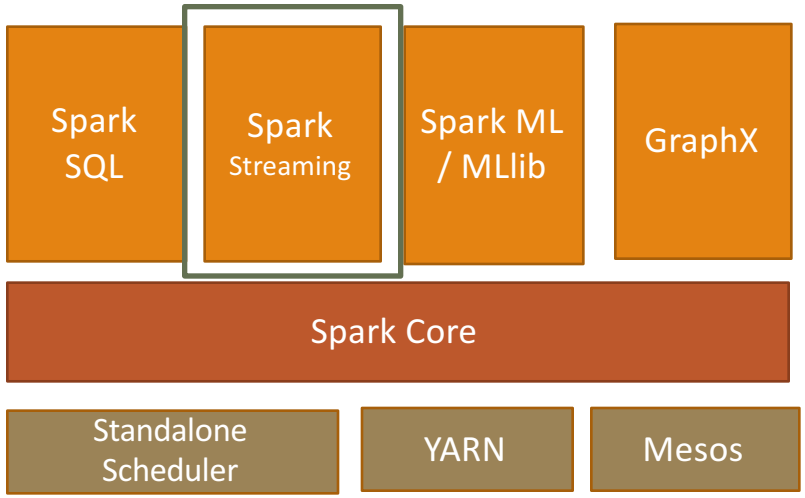
# Spark Interview Questions

- ~~What are the advantages of using Apache Spark over Hadoop MapReduce for big data processing?~~
- ~~What are the languages supported by Apache Spark for developing big data applications?~~
- ~~Can you use Spark to access and analyze data stored in Cassandra databases?~~
- ~~Is it possible to run Apache Spark on Apache Mesos?~~
- ~~How can you minimize data transfers when working with Spark?~~
- ~~Why is there a need for broadcast variables?~~
- ~~Name a few companies that use Apache Spark in production.~~
- ~~What are the various data sources available in SparkSQL?~~
- ~~What is the advantage of a Parquet file?~~
- ~~What do you understand by Pair RDD?~~
- ~~Is Apache Spark a good fit for Reinforcement learning?~~



https://www.dezyre.com/article/top-50-spark-interview-questions-and-answers-for-2016/208

# Data Stream Management with Spark



# Data Stream

# Data Stream Analysis

## Applications

- Social network
- Sensor data
- Online advertising
- Stock
- Website clickstreams
- IT logs
- Location-tracking
- Many more!

## Requirement

- Scalable
- Fault-torelent

# Spark Streaming

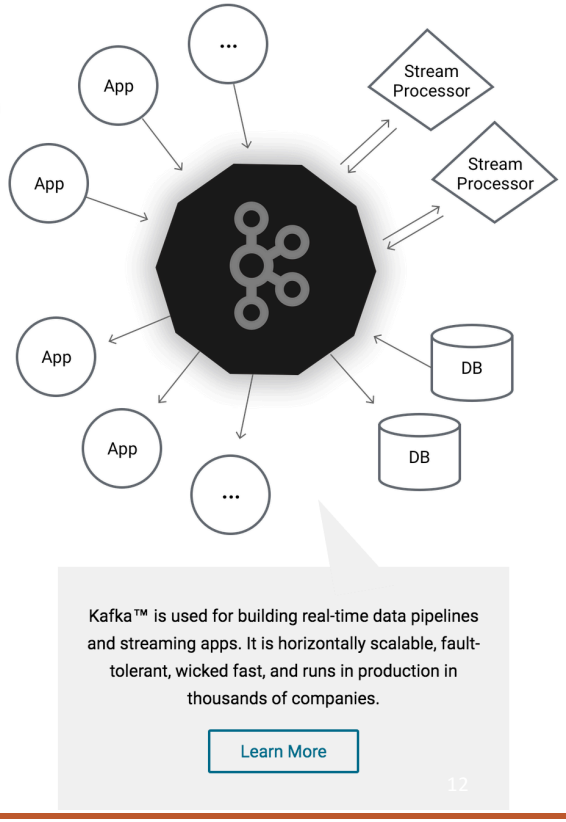
Read data from Hadoop compatible file systems (HDFS, S3, etc.) and distributed system.

Good performance with data streams. (Scalable and Fault tolerant.)

# Spark Streaming Data Sources

Apache Kafka :

- Build real-time data pipelines and streaming apps.
- Distributed, fast, scalable publish-subscribe messaging system.

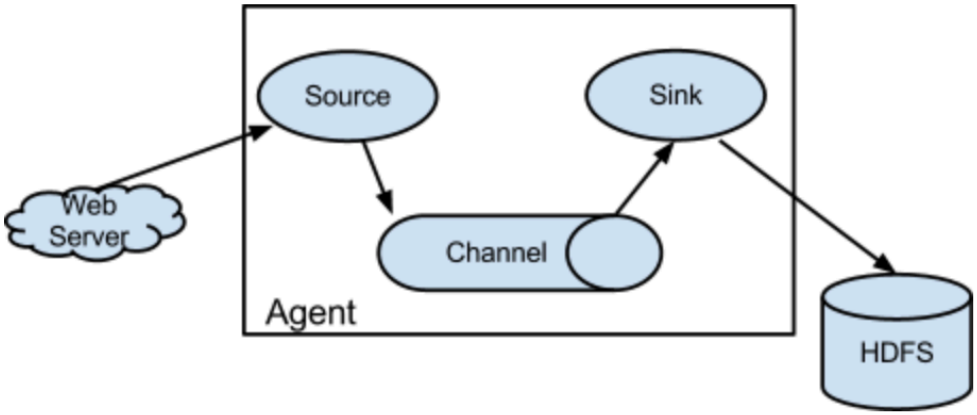




https://kafka.apache.org/intro.html

# Spark Streaming Data Sources

Apache Flume:

- Distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data.
- It has a simple and flexible architecture based on streaming data flows.


https://flume.apache.org/

# Spark Streaming Data Sources

Amazon Kinesis

- An AWS streaming platform.
- Build custom applications that process or analyze streaming data.


https://aws.amazon.com/kinesis/

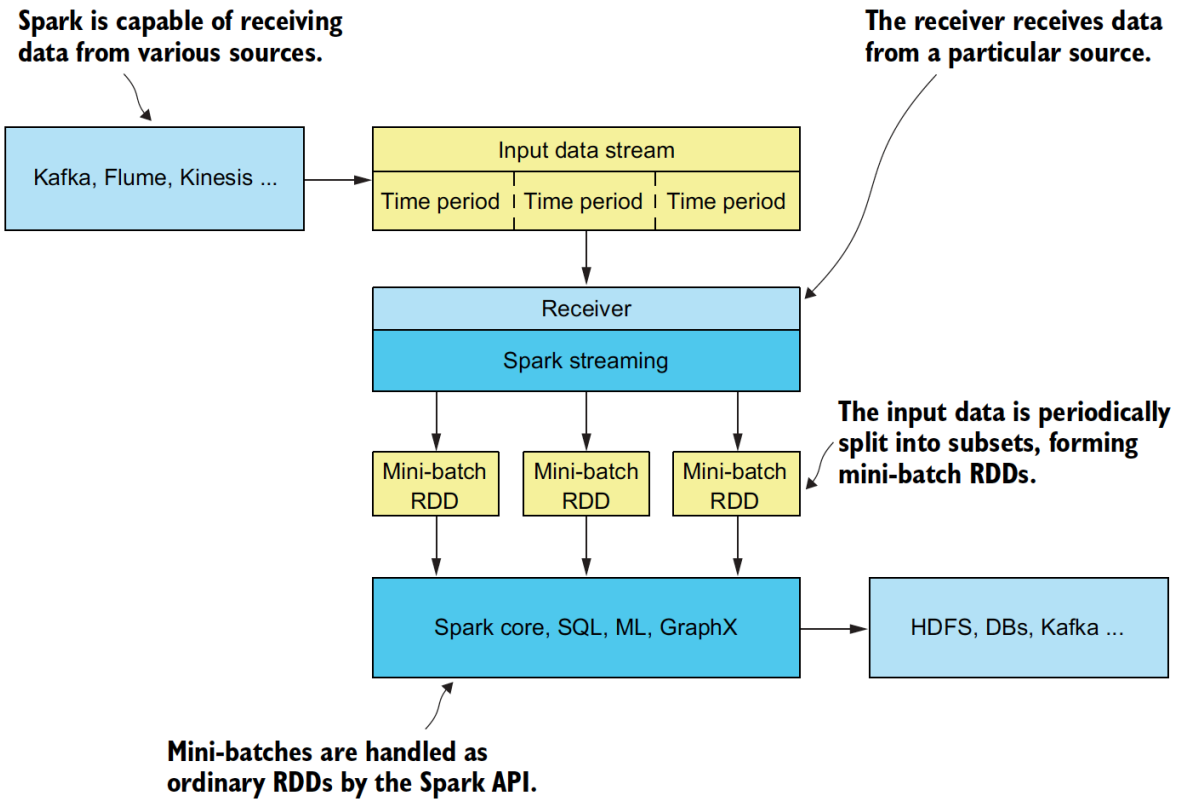
# Spark Streaming Data Sources

Twitter

- Provides APIs.


*https://dev.twitter.com/index*

# Spark Streaming



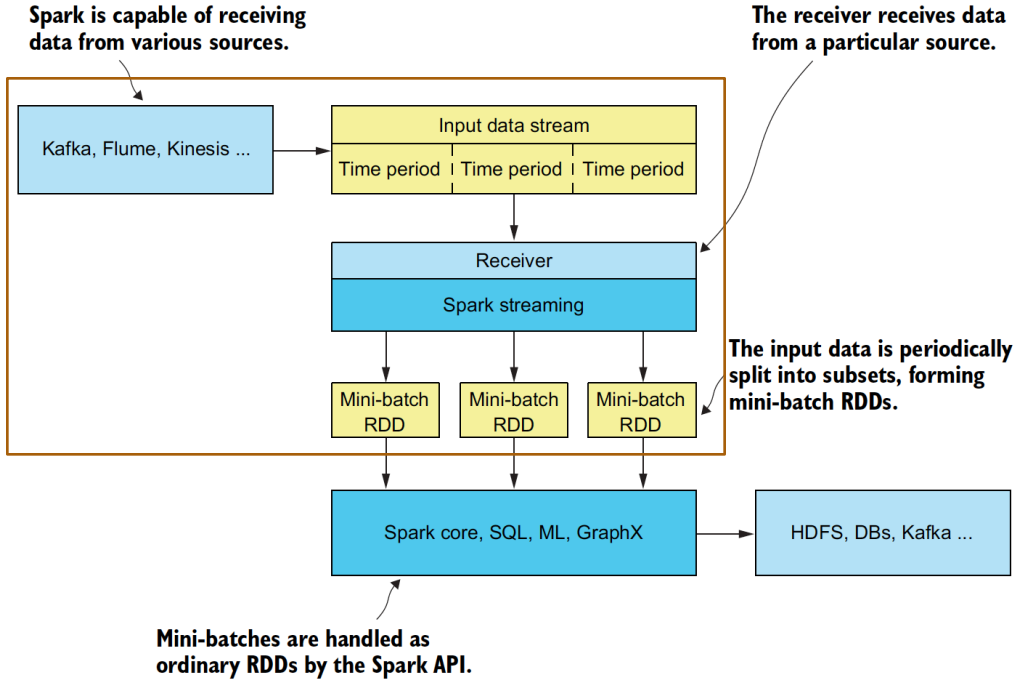
# Spark Streaming

## Spark Streaming

1. Receive blocks of data coming in specific time period
2. Split the incoming data into mini-batch RDDs.
3. Package as mini-batch RDDs.

Use other applications including Spark SQL, ML/MLlib, etc.

The result can be written to filesystems, relational DBs, or NoSQL.



# Summary


Extends the MapReduce model with primitives for efficient data sharing (Using Resilient Distributed Datasets (RDDs)).

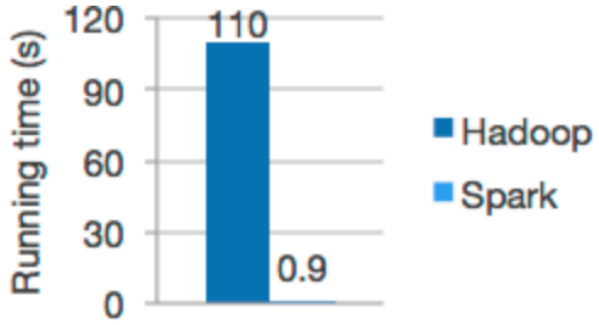
Achieves

- Speed.
- Ease of Use.
- Generality.
- Runs everywhere.


*Lightning-fast cluster computing*

Speed: Run programs up to 100x faster than Hadoop MapReduce in memory, or 10x faster on disk.

- Apache Spark has an advanced DAG execution engine that supports cyclic data flow and in-memory computing.




Logistic regression in Hadoop and Spark

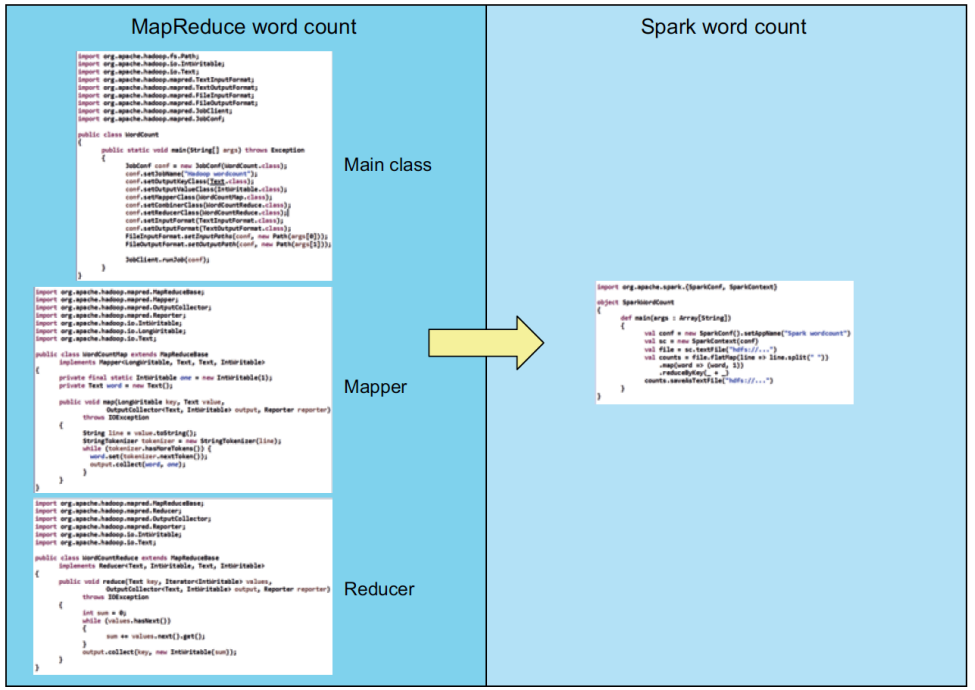

Ease of Use : Write applications quickly in Java, Scala, Python, R.

- Spark offers over 80 high-level operators that make it easy to build parallel apps. And you can use it interactively from the Scala, Python and R shells.

```
text_file = spark.textFile("hdfs://...")

text_file.flatMap(lambda line: line.split())
    .map(lambda word: (word, 1))
    .reduceByKey(lambda a, b: a+b)
```

Word count in Spark's Python API

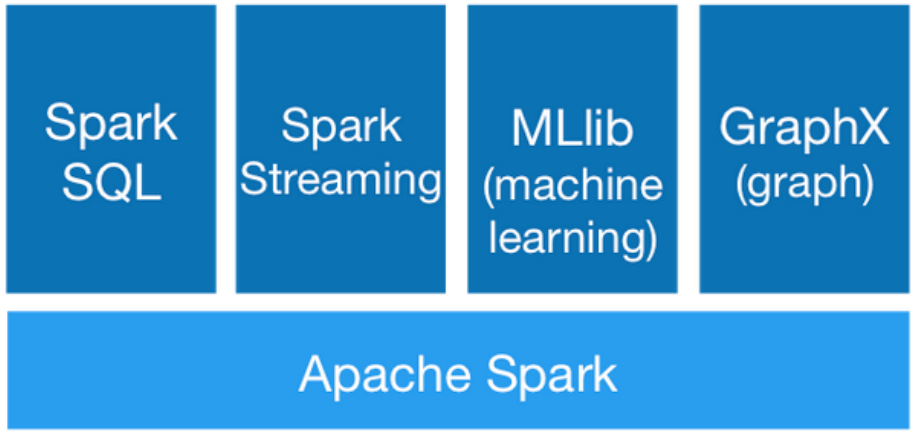



http://spark.apache.org/


*Lightning-fast cluster computing*

Generality : Combine SQL, streaming, and complex analytics.

- Spark powers a stack of libraries including SQL and DataFrames, Mllib and ML for machine learning, GraphX, and Spark Streaming. You can combine these libraries seamlessly in the same application.




http://spark.apache.org/


*Lightning-fast cluster computing*

Runs everywhere : Spark runs on Hadoop, Mesos, standalone, or in the cloud. It can access diverse data sources including HDFS, Cassandra, HBase, and S3.

- You can run Spark using its standalone cluster mode, on EC2, on Hadoop YARN, or on Apache Mesos. Access data in HDFS, Cassandra, HBase, Hive, Tachyon, and any Hadoop data source.

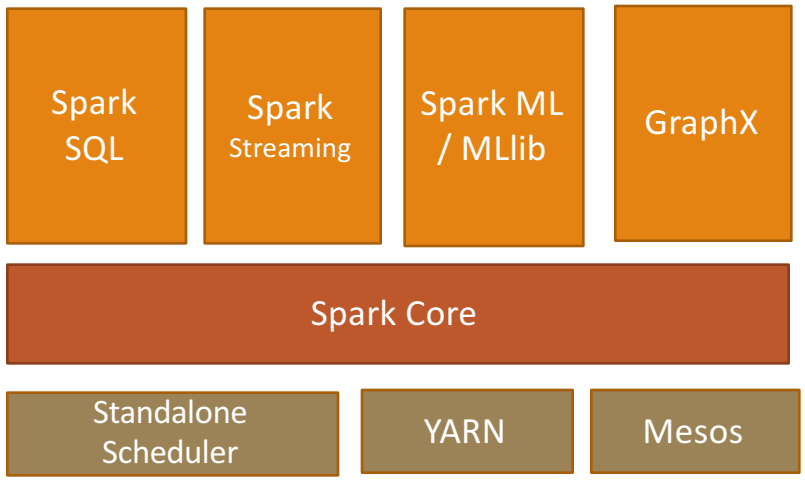




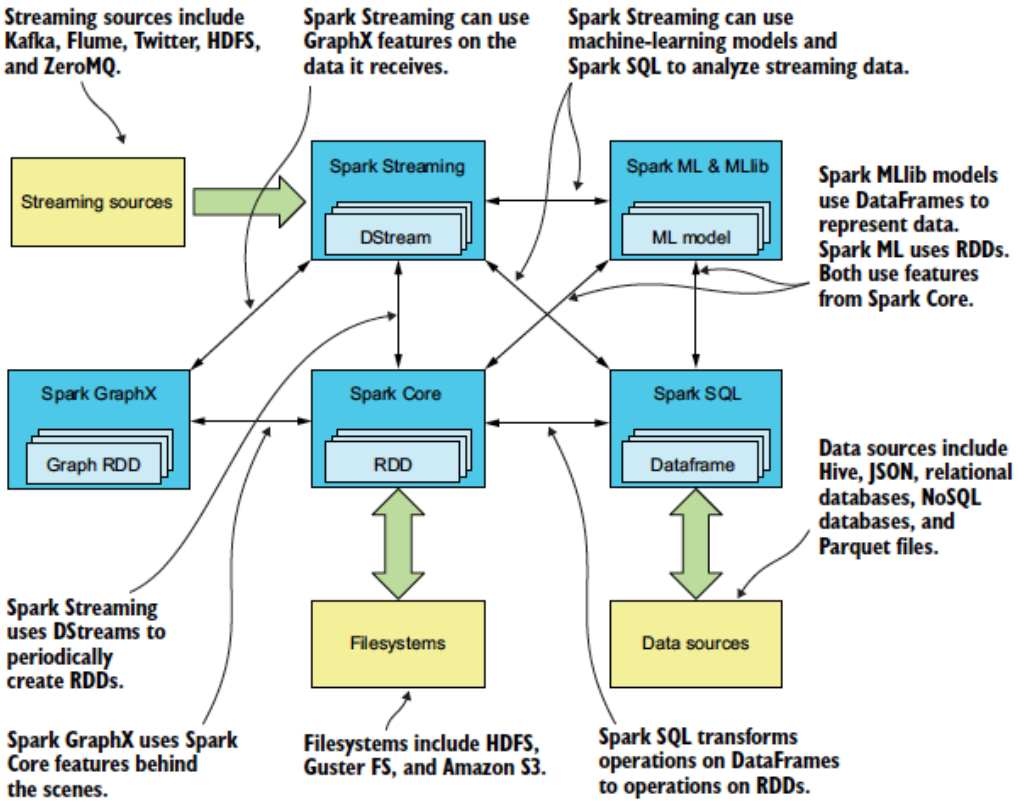

Community

- Most active open source community for big data.
- 200+ developers, 50+ companies contributing.

# Class Summary



# Class Summary



# Class Review On Canvas

# Reference

Spark Online Documentation : http://spark.apache.org/docs/latest/

Karau, Holden, et al. *Learning spark: lightning-fast big data analysis*. O'Reilly Media, Inc., 2015.

Zecevic, Petar, et al. Spark in Action, Manning, 2016.