.conf2015

Splunk and Spark

Liu-yuan Lai

Software Engineer, Splunk

splunk>

Disclaimer

During the course of this presentation, we may make forward looking statements regarding future events or the expected performance of the company. We caution you that such statements reflect our current expectations and estimates based on factors currently known to us and that actual events or results could differ materially. For important factors that may cause actual results to differ from those contained in our forward-looking statements, please review our filings with the SEC. The forward-looking statements made in the this presentation are being made as of the time and date of its live presentation. If reviewed after its live presentation, this presentation may not contain current or accurate information. We do not assume any obligation to update any forward looking statements we may make.

In addition, any information about our roadmap outlines our general product direction and is subject to change at any time without notice. It is for informational purposes only and shall not, be incorporated into any contract or other commitment. Splunk undertakes no obligation either to develop the features or functionality described or to include any such feature or functionality in a future release.

Agenda

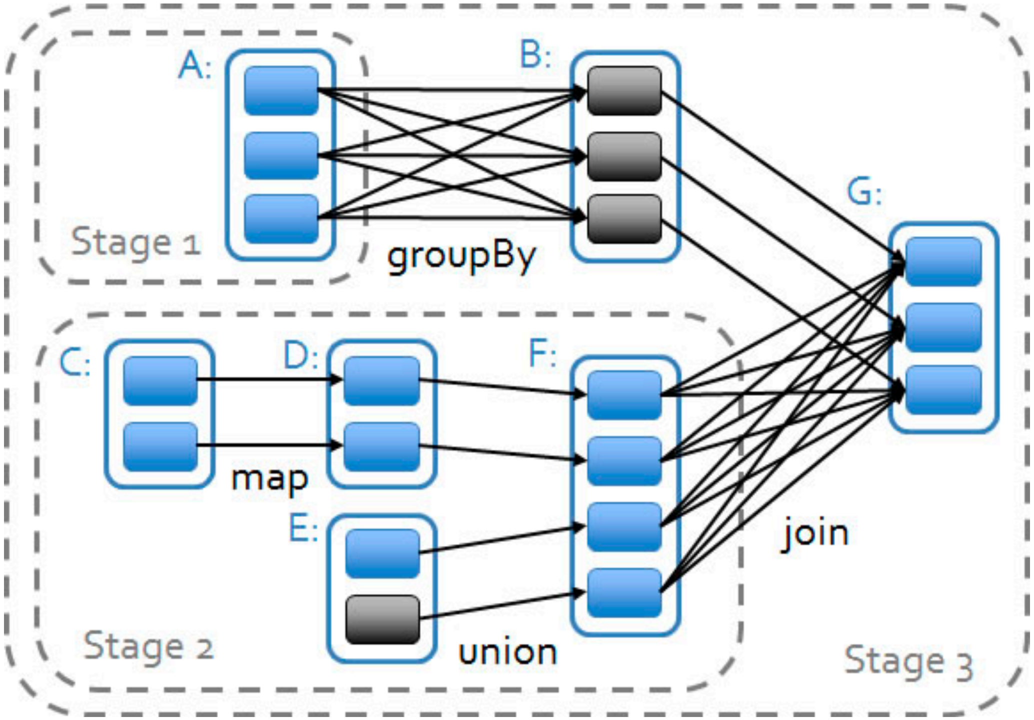
- Background: Spark
- Motivation
- Spark on Splunk – Splunk data as Spark external dataset
- Splunk with Spark – Extend Splunk search with Spark's computing power

Review: Spark basics

What is Spark?

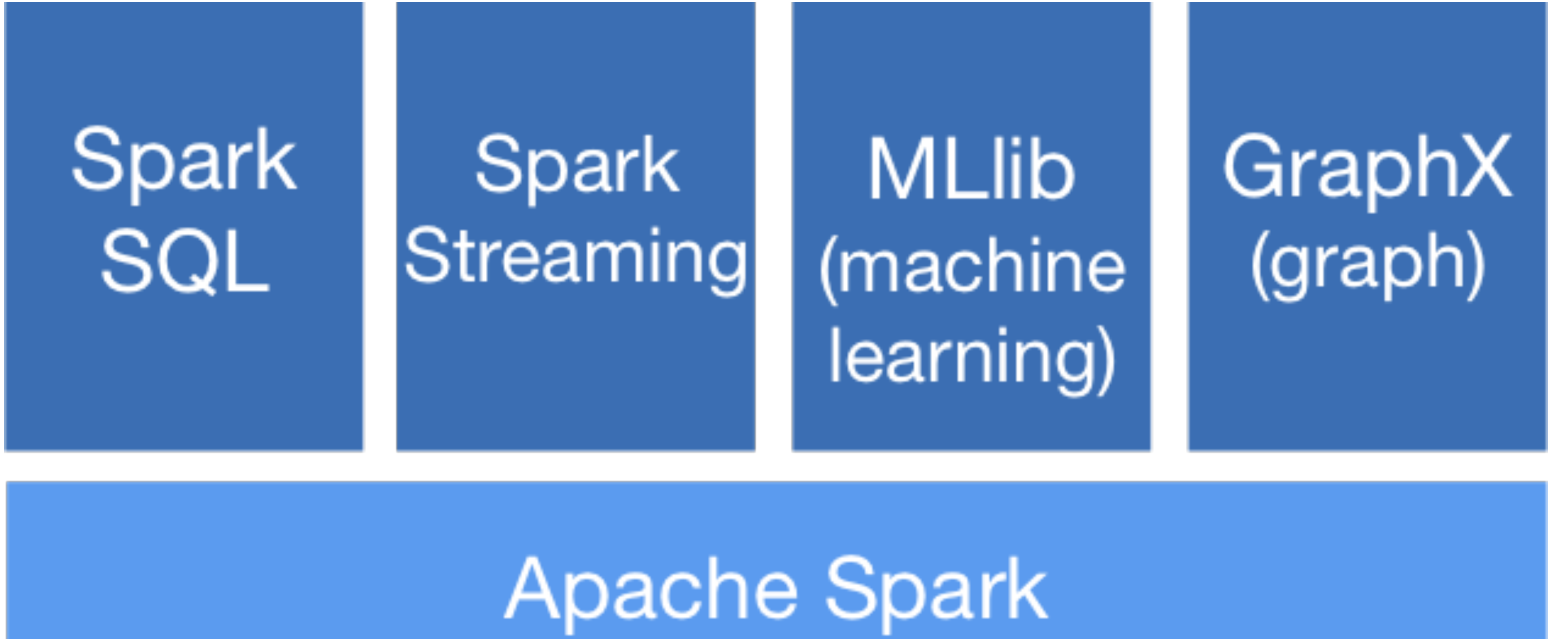
- “Apache Spark is a fast and general-purpose cluster computing system.”
- Abstract data as a distributed collection of RDD that can be operated in parallel.
- RDD have operations
  - Transformations: create a new dataset from an existing one
  - Actions: return a value after running a computation on the dataset.

Spark RDD



source: kmoses

Spark Stack



Spark SQL

Spark Streaming

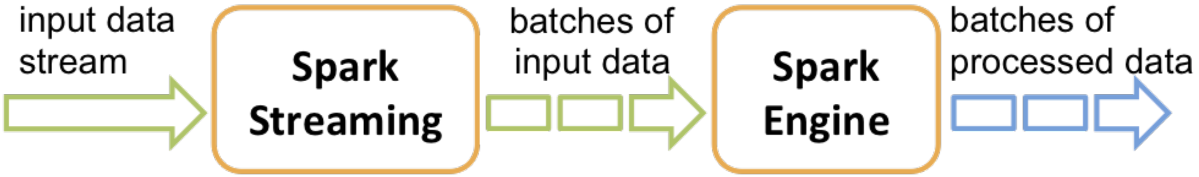
MLlib (machine learning)

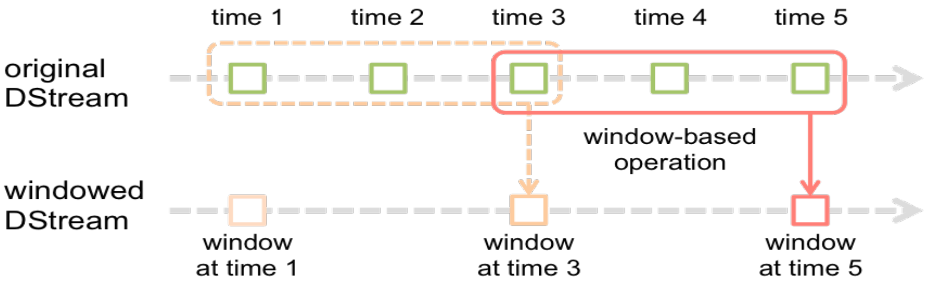
GraphX (graph)

Apache Spark

source: apache spark

Spark Streaming





source: apache spark

Spark on Splunk

What Splunk Brings To Spark

Splunk

- Great support on data ingestion
- Unstructured/semi-structure data indexing
- Powerful runtime data 'wrangling' through splunk search language
  - eventtypes
  - fields extractions
  - tags
  - lookups

Spark supported datasets

- Local filesystem
- Hadoop HDFS
- Cassandra
- Hbase
- Amazon S3

Spark on Splunk

- SplunkRDD
  - RDD from splunk search results
- SplunkDStream
  - DStream from splunk realtime search
- SplunkUtils
  - createRDD
  - createStream

Demo

Buttercup Games *(limited in preview, will enrich the content)*

Splunk with Spark

Beyond Data Processing

- Ability to perform analysis and machine learning on data
- Challenges:
  - Algorithms could be complex and not expressible by splunk search language
    - Custom search commands
  - Wide variety of models and algorithms
    - Search commands overload
  - Dataset-dependent
    - Search command parameters overload
  - Repeated trial, training, testing and fine-tuning

Extend Splunk with Spark

- Distributed computing for complex operations
- Impressive arsenal by Spark stack is readily available
- Users write their own spark programs tailored to their dataset
- Connect other data sources, through RDD/DStream

Problem Statement

- Study data
- Experiments
  - Select algorithms/models
  - train
  - Test
  - Validate
- Apply to production
- Monitor and fine-tune

Design Goal

Make it natural to perform analysis and learning in splunk

- ☑Study data
- Experiments
- Production
- *Monitor
- *Fine-tune

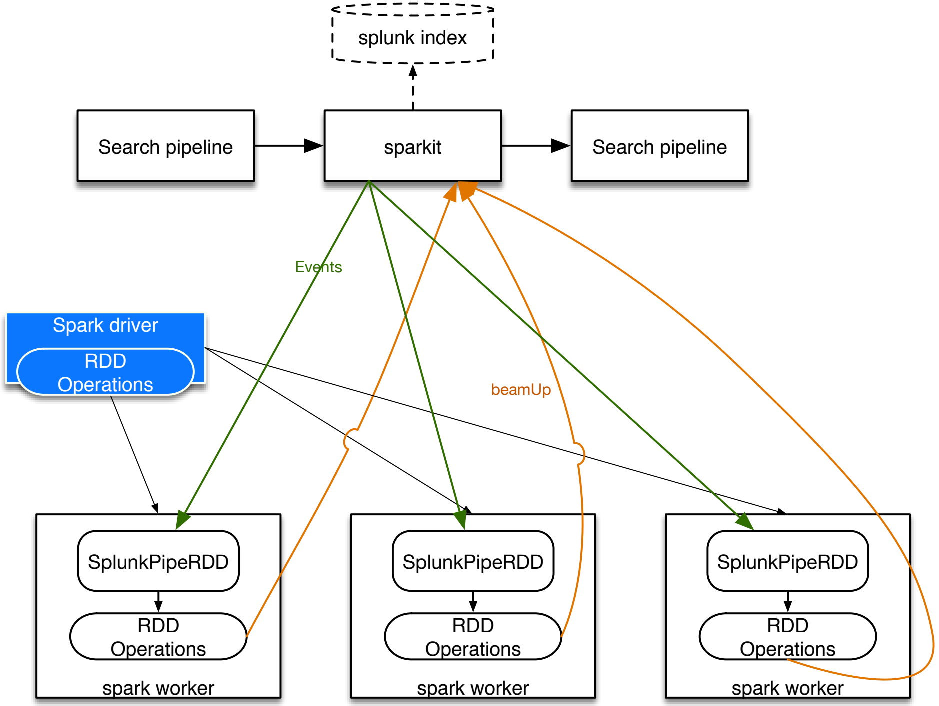
Design Choices

- Integrate spark into splunk search language
  - Spark becomes an extended context of Splunk for complex computations
- Simple command interface
- Do not impose limitations on operations that can be run on Splunk events
- Do not run user codes within splunk process
- Interactive inspection and tuning
- Splunk for ML 'Experiment' management

Approach

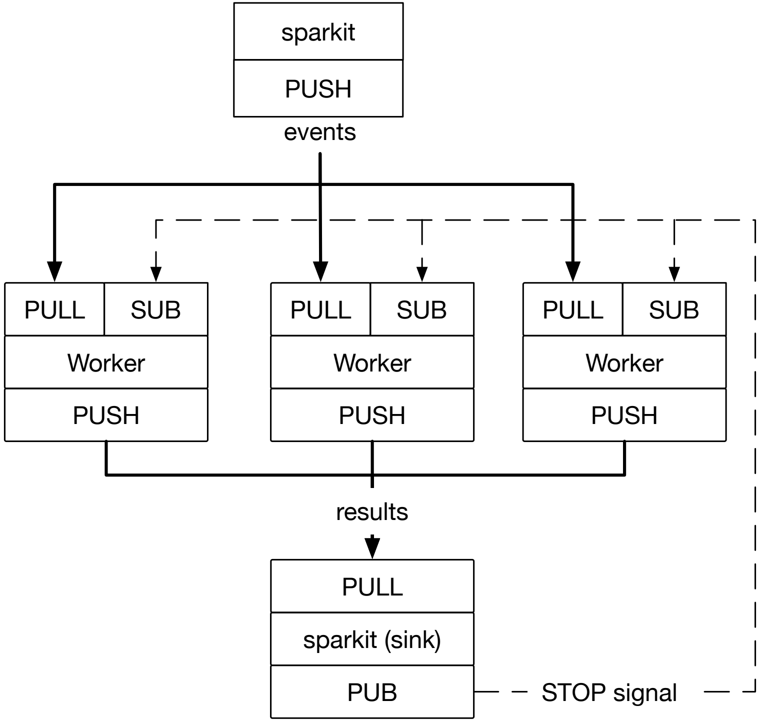
- Search command “sparkit”
  - starting point to distribute search pipeline results to spark
- SplunkPipeRDD
  - RDD that pulls data from search pipeline (sparkit)
- Custom RDD operation “beamUp”
  - Push computation results to search pipeline (sparkit)

Architecture



Implementation

ØMQ for communicating data and results



Demo

- *(limited in preview, will enrich before .conf)*

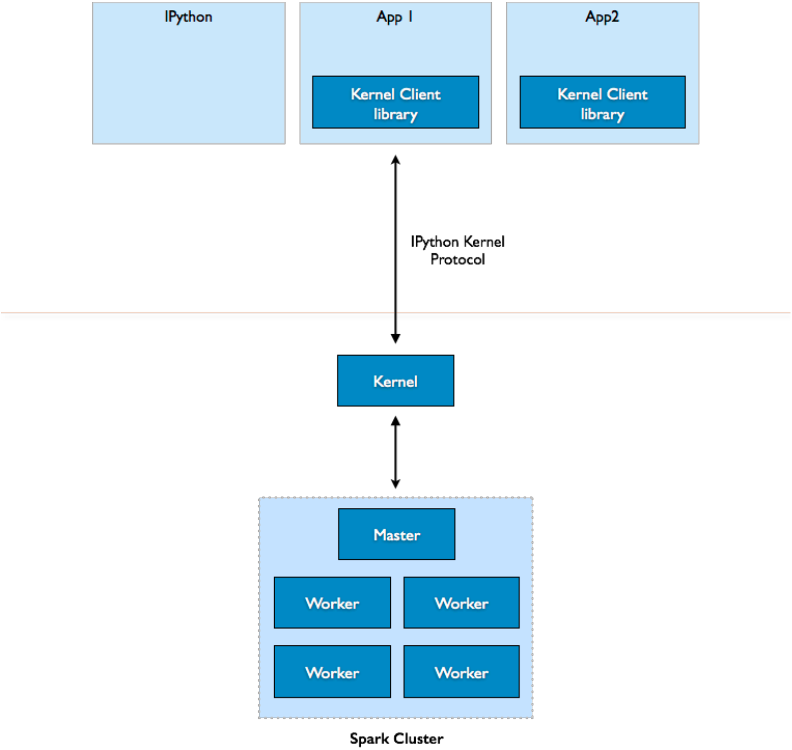
Future Works - Features

- Splunk for ML ‘Experiment’ management
  - Index ‘sparkit’ results into a separate index, each run given some sort of id
- Use IPython kernel approach (spark kernel) for better flow control
- DataFrame
- ML model or “computation” registration

ML Management

- Write results to ‘experiment’ index
  - Enables tapping in realtime or post-mortem
- Search by job-id to retrieve testing/training/production results for further investigation
- Add metadata for experiments management

Spark Kernel



source: [spark kernel project]()

DataFrame and SparkSQL

- “Spark SQL is a Spark module for structured data processing. It provides a programming abstraction called DataFrames and can also act as distributed SQL query engine.”
- DataFrame: SQL/table-like query and operations; enables many new optimizations starting Spark 1.5

```
df.select(df("name"), df("age") + 1).show()
// name    (age + 1)
// Michael null
// Andy    31
// Justin  20
```

- SplunkRDD automatically create schema from extracted fields

Future Works - Technical

- SplunkContext(?)
  - auto-discover mundane settings
- Splunk indexer and search-head clustering environment
- Performance, scalability
- Fault tolerance, stability

THANK YOU