Spark
Summit
-1,164 participants from over 453 companies attended
-Spark Training sold out at 300 participants
-31 organizations sponsored the event
-12 keynotes and 52 community presentations were given
News
-Project
-Databricks
Project
- 1.0.0 release
- Graduated incubator
- Very active community
Very active community
- Top three Apache projects
- Most active Big Data project
- 50 companies
- 250 contributors
- 175,000 LOC
Databricks
-Certification
-Cloud
Certification
- Every certified app will run on every certified distribution
- Distribution Partners
- App Partners
Distribution Partners
- Cloudera
- MapR
- Hortonworks
- Pivotal
- IBM
- Amazon Web Services
- SAP
App Partners
- Alteryx
- Datastax
- 0xdata
- Typesafe
- Zoomdata
Cloud
-Vision: Make Big Data Easy!
-Product: Badass
-Hosted Platform
-Cluster Management
-Interactive Workspace
Interactive Workspace
- Notebooks
- Dashboards
- Jobs
Dashboards
- WYSIWYG Builder
- Interactive plots
- One-click publishing
Spark Basics
- Execution
- RDDs
- Caching
- Broadcast
- Languages
Execution
- Apply Functional Operators across Distributed Collections
- Master / Worker
- Lazy
- Parallelize with Threads first
RDDs
- Interface for dataset
- Backed by anything
- Any InputFormat class
- HDFS default
Caching
-Store intermediate results in memory
-Partition-locality
-Significant speed-up for iterative algorithms
Broadcast
- Send immutable object to all workers
- Similar to DistributedCache in mapreduce
Languages
-Scala
-Python
-Java 7
-Java 8
-R
-Clojure
Advanced
- Partitioning
- Persistence Options
- Checkpointing
- Accumulators
- Optimizations
Subprojects
- SparkSQL
- Tachyon
- Spark Streaming
- MLLib
- GraphX
- BlinkDB
- Spark Job Server
SparkSQL
- Replaces Shark Core Catalyst Libraries
Core
-SchemaRDDs
-Query Execution
-Caching
Catalyst
- Relational algebra
- Expressions / UDFs
- Query Planning
- Optimizer
Libraries
-POJOs
-JDBC
-JSON
-Parquet
-Hive
Hive
- Catalog info from Metastore
- Helps connect UI like Microstrategy / Tableau
- Wrappers for UDF, UDAFs, UDTFs
- Supports TRANSFORM
- Supports SerDes
Tachyon
- In Memory (Off-Heap) Distributed Datastore
- Change URI from hdfs:// to tachyon://
- Share datasets between jobs without HDFS
- Helps scaling by off-loading allocation responsibility and GC pauses from executor processes
Spark Streaming
- Real-time streams
- Micro-batching
- Windowed Computations
- Lambda Architecture
MLLib
- Summary statistics
- Regression
- Classification
- Clustering
- Collaborative Filtering
- Optimization
- Dimensional Reduction
GraphX
- Graph, VertexRDD, EdgeRDD objects and operations
- Pregel API
- mapReduceTriplets List<V,E,V>
- Graph analytics libraries
Graph analytics libraries
- ConnectedComponents
- PageRank
- TriangleCount
- ShortestPaths
- SVDPlusPlus
BlinkDB
- Get estimated results
- Time bound
- Error bound
Spark Job Server
- Runs multiple jobs / contexts in same process
- Allows for RDD Caching / Sharing between jobs
- Job Persistence
Use Cases
- Spotify
- Real-time Auctions - ShareThrough
- Real-time Recommendations - Graphflow
- Cancer Genomics - AMPLab
- Malware Detection - F-Secure
- Media Distribution Analytics - NBC Universal
- Personal Fitness - Jawbone
- Neuroscience - HHMI
Resources
- Code
- Event
- Technology
- Videos
Code
-https://github.com/apache/spark
Event
- spark-summit.org
- http://arjon.es/2014/06/30/spark-summit-2014-day-1/
- https://www.crowdchat.net/chat/c3BvdF9vYmpfODc=.
- https://nathanbrixius.wordpress.com/2014/07/02/spark-summit-keynote-notes/
- http://thomaswdinsmore.com/2014/07/03/spark-summit-2014-roundup/
Technology
-Learning Spark (O'Reilly eBook)
-www.spark-stack.org
-ampcamp.berkeley.edu
YouTube
-AmpLab  
https://www.youtube.com/channel/UCWudC4d9i-2yxR5tuen-Nuw

-Databricks  
https://www.youtube.com/channel/UC3q8O3Bh2Le8Rj1-Q-_UUbA

-Apache Spark  
https://www.youtube.com/channel/UCRzsq7k4-kT-h3TDUBQ82-w