1. Velocity

Spark Releases

Spark Jira Tickets
Spark Releases
Spark Jira Tickets
2. Keeping Up

Mailing Lists
user@spark.apache.org
dev@spark.apache.org

Jira

My Blog
data-science-as-a-service.org/blogs/michaelbrinke

Conference Videos

Code Repositories
https://github.com/amplab
https://github.com/amplab-extras
http://spark-packages.org/

Google Scholar Alerts
Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing

Twitter

Books
Mailing Lists

user@spark.apache.org
dev@spark.apache.org
Jira

[Webpage screenshot of Jira issue tracking system with a specific issue: SPARK-6084, titled "spark-shell broken on Windows". The issue is described with details about the problem, steps to reproduce, and a description of the error message.]
My Blog

datascienceassn.org/blogs/michaelmalak
Conference Videos

ampcamp
Big Data Bootcamp

Spark Summit

Bay Area Spark Meetup

Spark SQL Data Sources: overview of API & HBase data source from Huawei

Strata+Hadoop World
Books
Google Scholar Alerts

Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing
Code Repositories

https://github.com/amplab
https://github.com/amplab-extras
http://spark-packages.org/
3. Ecosystem

Add-Ons

- AMP Lab
- Apache Hive
- Spark Packages.org

Spark

- Core
- Spark SQL
- Spark Streaming
- MLlib
- GraphX

Dependencies

- Distributed Storage
- Cluster Manager

- Hadoop
- YARN
- Mesos
- Core
- Spark SQL
- Spark Streaming
- MLLib
- GraphX
Dependencies

Distributed Storage

Cluster Manager

- Standalone
- YARN
- Mesos
Distributed Storage
Cluster Manager

- Standalone
- YARN
- Mesos
Add-Ons

**AMP Lab**
- Almost in Apache Spark
  - Spark 3.0
  - Managed Ecosystem
  - Spark on AWS
- Machine Learning
  - AI
  - Machine Learning

**Core Technology**
- Tez
- Tachyon
- HDFS

**Applications**
- ADAM
- Cask

---

**Apache Hive**
HIVE-7292: Hive-on-Spark

---

**Spark-Packages.org**
- Spork
- HBase connectors (3)
- Sparkling Water
- Spark-MongoDB
- Thunder/Lightning
AMP Lab

Almost in Apache Spark
- SparkR
- IndexedRDD
- BlinkDB

Core Technology
- Tachyon
- Succinct
- Sparrow

Machine Learning
- MLBase
- ML Optimizer
- MLI
- ML Matrix
- CoCoA
- Velox

Applications
- ADAM
- Carat
Almost in Apache Spark

- SparkR
- IndexedRDD
- BlinkDB
Core Technology

- Tachyon
- Succinct
- Sparrow
Applications

- ADAM
- Carat
Machine Learning

- MLBase
  - ML Optimizer
  - MLI
- ML Matrix
- CoCoA
- Velox
Apache Hive

HIVE-7292: Hive-on-Spark

https://issues.apache.org/jira/browse/HIVE-7292

Hive / HIVE-7292

Hive on Spark

Details
Type: Improvement
Priority: Major
Affects Version/s: None
Component/s: Spark

Status: OPEN
Resolution: Unresolved
Fix Version/s: None

Description
Spark as an open-source data analytics cluster computing framework has gained significant momentum recently. Many Hive users already have Spark installed as their computing backbone. To take advantages of Hive, they still need to have either MapReduce or Tez on their cluster. This initiative will provide user a new alternative so that those user can consolidate their backend.

Secondly, providing such an alternative further increases Hive's adoption as it exposes Spark users to a viable, feature-rich de facto standard SQL tools on Hadoop.

Finally, allowing Hive to run on Spark also has performance benefits. Hive queries, especially those involving multiple reduce stages, will run faster, thus improving user experience as Tez does.

This is an umbrella JIRA which will cover many coming subtask. Design doc will be updated.
Spark-Packages.org

- Spork
- HBase connectors (3)
- Sparkling Water
- Spark-MongoDB
- Thunder/Lightning