State of Spark, and where it is going

Reynold Xin @rxin
Strata Singapore
Dec 3rd, 2015
Spark stack diagram

SQL  Streaming  MLlib  GraphX

Spark Core (RDD)
A Great Year for Spark

Most active open source project in big data

New language: R

Widespread industry support & adoption
Community Growth

Summit Attendees

<table>
<thead>
<tr>
<th>Year</th>
<th>2014</th>
<th>2015</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1100</td>
<td>3900</td>
</tr>
</tbody>
</table>

Meetup Members

<table>
<thead>
<tr>
<th>Year</th>
<th>2014</th>
<th>2015</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>12K</td>
<td>50K</td>
</tr>
</tbody>
</table>

Developers Contributing

<table>
<thead>
<tr>
<th>Year</th>
<th>2014</th>
<th>2015</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>500</td>
<td>1000</td>
</tr>
</tbody>
</table>
Meetup Groups: December 2014
Open Source Ecosystem

Applications
- Sparkling
- H2O
- IP
- Apache Ambari
- thunder
- oqoop

Environments
- Hadoop
- kubernetes
- MESOS
- docker
- spring
- openstack

Data Sources
- MySQL
- PostgreSQL
- cassandra
- HDFS
- mongoDB
- Parquet
- TACHYON
- elasticsearch
- HBase
- SequoiaDB
- Kafka
Users
1000+ companies

Distributors + Apps
50+ companies
Diverse Runtime Environments

Cluster Managers

- 48% Standalone mode
- 40% YARN
- 11% Mesos

51% on a public cloud
Industries Using Spark

- **29.4%**
  Software (SaaS, Web, Mobile)

- **14.0%**
  Consulting (IT)

- **9.6%**
  Advertising, Marketing, PR

- **17.7%**
  Other

- **6.5%**
  Banking, Finance

- **4.4%**
  Computers, Hardware

- **4.4%**
  Education

- **6.7%**
  Retail, e-Commerce

- **3.9%**
  Health, Medical, Pharmacy, Biotech

- **3.5%**
  Carriers, Telecommunications
Top Applications

- Business Intelligence: 68%
- Data Warehousing: 52%
- Recommendation: 44%
- Log Processing: 40%
- User-Facing Services: 36%
- Fraud Detection / Security: 29%
Largest Cluster & Daily Intake

- 800 million+ active users
- 8000+ nodes
- 150 PB+ 1 PB+/day
Alibaba Taobao

- clustering (community detection)
- belief propagation (influence & credibility)
- collaborative filtering (recommendation)

* Spark Summit San Francisco 2014
Top Retail Bank & Huawei

Structured Data

DB/DW

Huawei FusionInsight Spark

Structured, Semi-Structured, Unstructured Data

Credit Proof

2~5 Seconds

Credit Proof

2~5s

On Line History query

7 years+

History Query

1 year

Micro-loan Conversion Rate

Higher

40X

Structured, Semi-Structured, Unstructured Data

Possible Assets

Financial Networking

Targeted Marketing

Top Retail Bank & Huawei
Are We Done?

No! Development is faster than ever. Expect Spark 2.0 in 2016.

Biggest technical change in 2015 was DataFrames
  • Moves many computations onto the relational Spark SQL optimizer

Enables both new APIs and more optimization, which is now happening through Project Tungsten
Coming in Spark 1.6

Dataset API: typed interface over DataFrames / Tungsten

- Common ask from developers who saw DataFrames

```scala
case class Person(name: String, age: Int)

val dataframe = read.json("people.json")
val ds: Dataset[Person] = dataframe.as[Person]

ds.filter(p => p.name.startsWith("M"))
  .groupBy("name")
  .avg("age")
```
Other Upcoming Features

DataFrame integration with GraphX and Streaming

More Tungsten features: faster in-memory cache, SSD storage, better code generation

Data sources for Streaming
Thank you.

@rxin