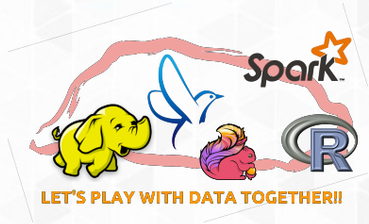
亦思科技
is-land Systems inc.



HadoopCon2016

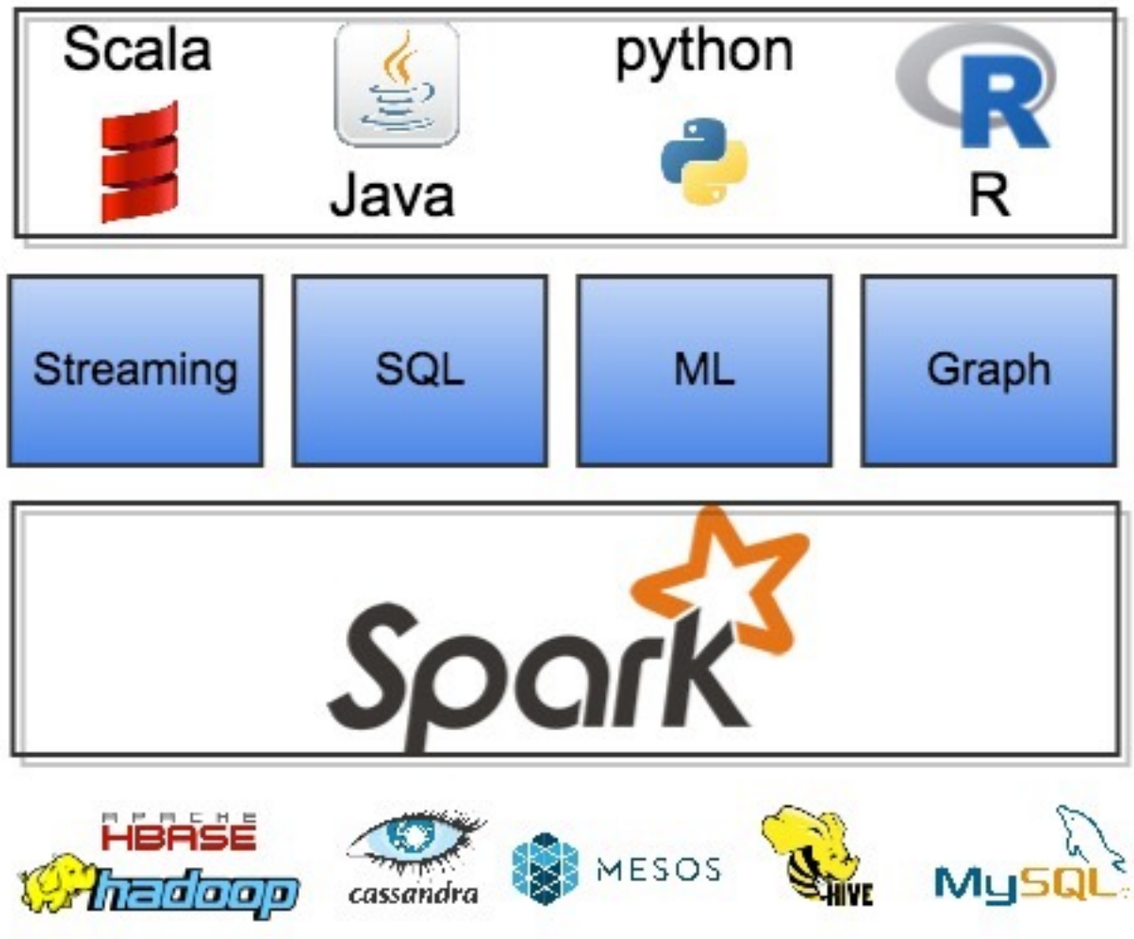
Spark Today

Stana 何永安
stana@is-land.com.tw

Who am I ?

- Stana 何永安
- Worked at is-Land
- Be familiar with Hadoop、HBase、Hive and Spark

Benefits of Apache Spark



- Speed
  - 100x faster than Hadoop for large scale data processing.
- Ease of Use
  - Easy-to-use APIs.
- Unified Engine
  - Packaged with higher-level libraries,including streaming data,SQL queries,machine learning and graph processing.

What's New in 2.0 ?

- **Structured API improvements**
  - DataFrames, Datasets, SparkSession

- **Structured Streaming**

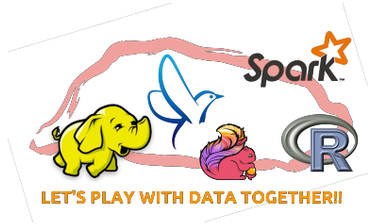
- MLlib model export

- MLlib expand Python、R APIs

- SQL 2003 support

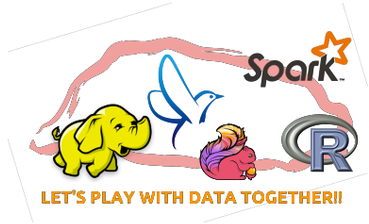
- Scala 2.12 support

Others



HadoopCon2016

- Experience with Apache Spark on video games



HadoopCon2016

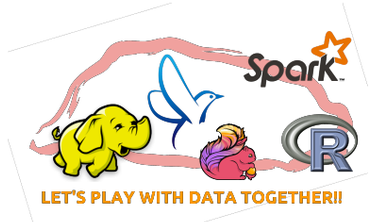
Structured API Improvements

What is Structure ?

- In DataFrame/Dataset, structure means that data is organized into named columns like a table in relational database.

- Metadata.

Why Structure?

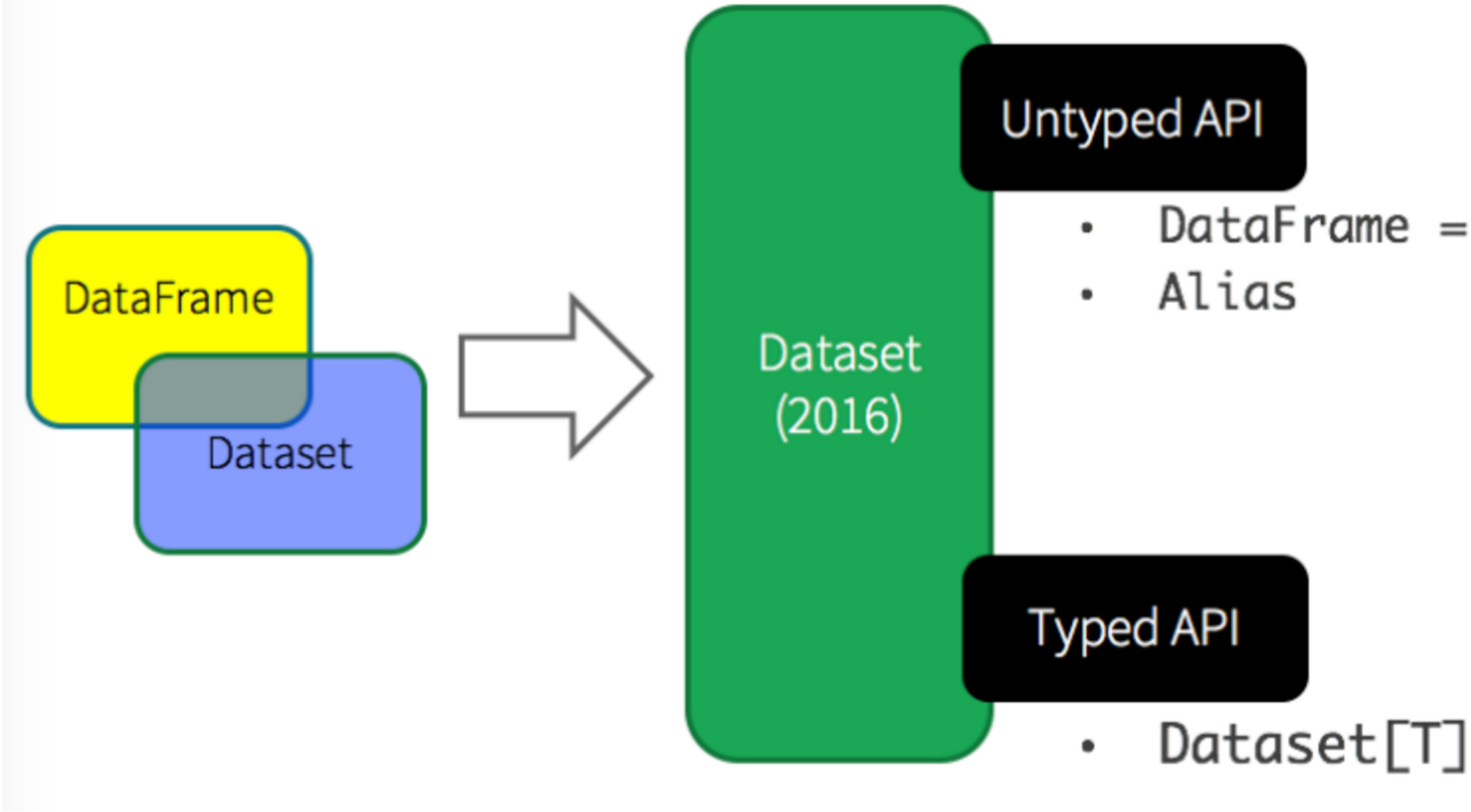


- Structure will limit what can be expressed, but enable optimizations

DataFrame

- Like a RDD, it is a distributed collection of data.
- Unlike a RDD, data is organized into named columns like a table in relational database.
- in Spark 2.0, DataFrame APIs will merge with Datasets APIs.
- DataFrame = DataSet[Row]
  - A data set of generic row objects.

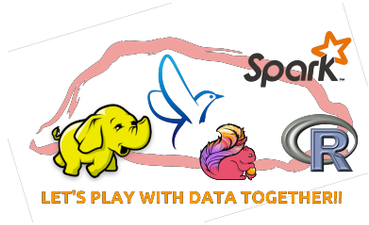
Unified Spark 2.0 API



reference:https://databricks.com/wp-content/uploads/2016/06/Unified-Apache-Spark-2.0-API-1.png

Dataset

- An extension of the DataFrame API
- Like a DataFrame, take advantage of performance.
- Compile-time type safety.
  - Applications can be checked for errors before they are run.
- Providing many of the same functional transformations(e.g. map, flatMap, filter) with RDD

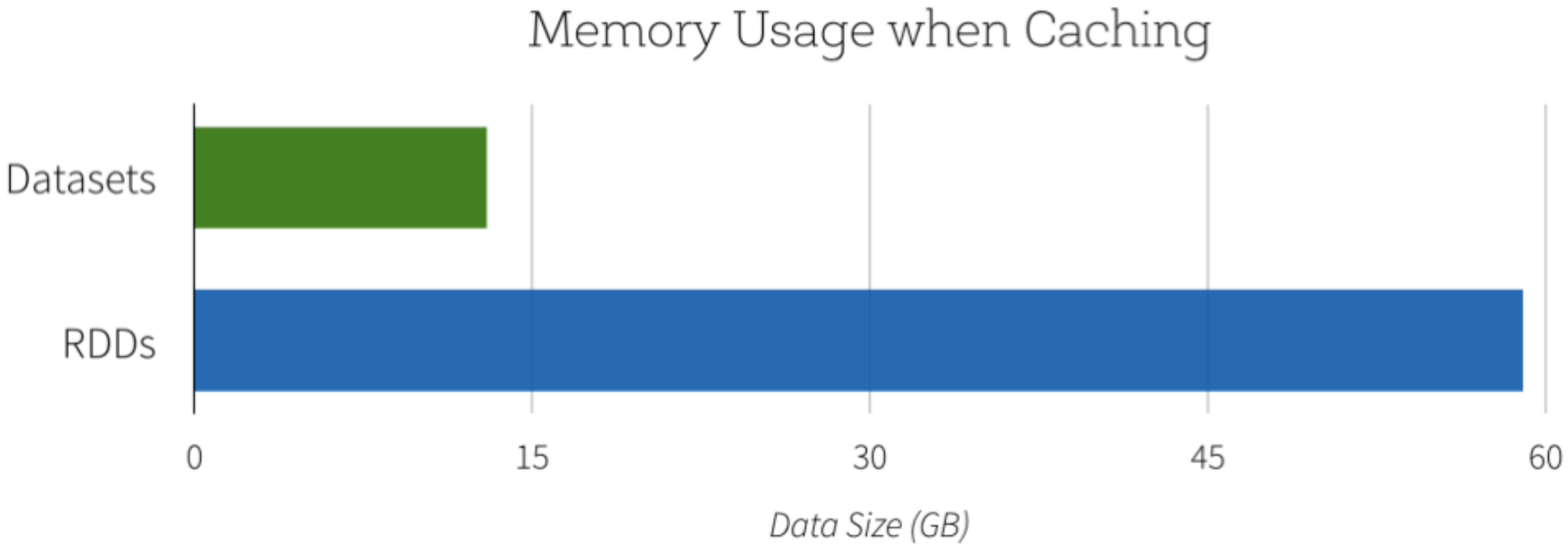


HadoopCon2016

APIs

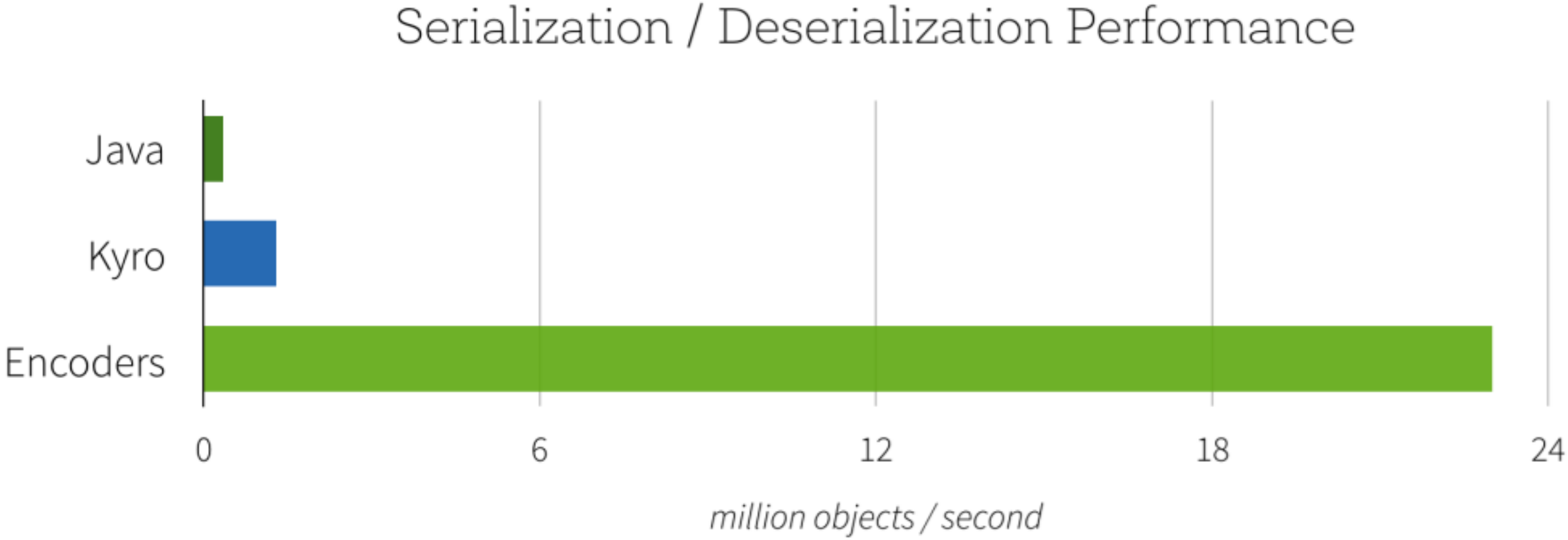
```
val df = spark.read.json("store.json")
case class Store(name: String, sales: Int)
// convert DataFrame to Dataset. (DataFrame = DataSet[Row])
val ds: Dataset[Store] = df.as[Store]
ds.filter(_.sales > 25000)  .show()
// +------+--------+
// | name |  sales  |
// +------+--------+
// |   null | 30120 |
// |  Andy| 28888 |
// | Justin| 40101 |
// +------+--------+
```

Space Efficiency

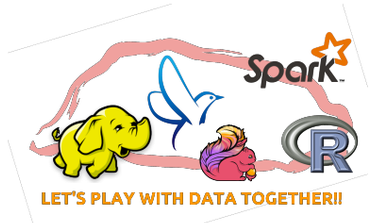


reference:http://www.slideshare.net/databricks/structuring-spark-dataframes-datasets-and-streaming-62871797

Serialization performance



reference:http://www.slideshare.net/databricks/structuring-spark-dataframes-datasets-and-streaming-62871797

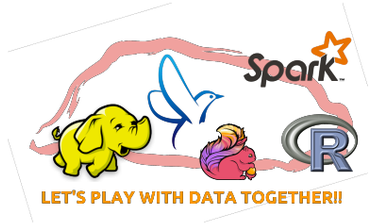


HadoopCon2016

When to use RDD, DataFrame, DataSet ?

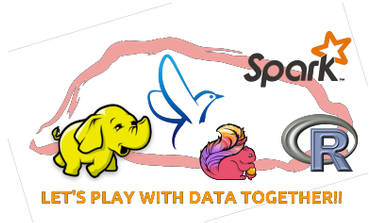
When to use RDDs?

- Wanting low-level transformation and actions and control on data.
- Data is unstructured, such as media streams.
- Giving up some optimization and performance benefits available with DataFrame/Dataset.



When should I use DataFrames or Datasets?

- If you want high-level abstractions APIs, use DataFrame or Dataset.
- If your processing expressions, filters, maps, aggregation, SQL queries, on structured data, use DataFrame or Dataset.
- If you want to take advantage of Catalyst optimization, and benefit from Tungsten's efficient code generation, use DataFrame or Dataset.
- If you want unification and simplification of APIs across Spark Libraries, use DataFrame or Dataset.



HadoopCon2016

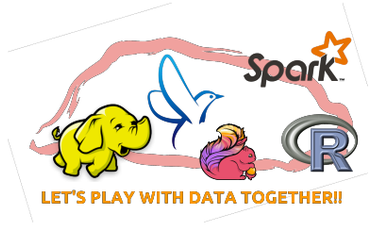
Unification and simplification

RDDs

```
val counts = words
    .groupBy(_.toLowerCase)
    .map(w => (w._1, w._2.size))
```

Dataset

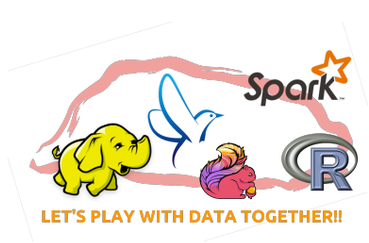
```
val counts = words
    .groupBy(_.toLowerCase)
    .count()
```



SparkSession

- A new entry point that combination SQLContext, HiveContext and future StreamingContext.

- In Spark 2.0, a new entry point for DataSet and Dataframe.



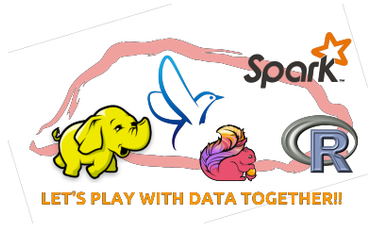
APIs

Spark 1.x

```
//set up the spark configuration
val sparkConf = new SparkConf().setAppName("SparkSessionExample").setMaster("local")
//create SparkContext to access SQLContext
val sc = new SparkContext(sparkConf).set("spark.config.option", "value")
val sqlContext = new org.apache.spark.sql.SQLContext(sc)
```

Spark 2.0

```
// Create a SparkSession. No need to create SparkContext
// You automatically get it as part of the SparkSession
val warehouseLocation = “/temp/spark-warehouse"
val spark = SparkSession
   .builder()
   .appName("SparkSessionExample")
   .config("spark.sql.warehouse.dir", warehouseLocation)
   .enableHiveSupport()
   .getOrCreate()
```

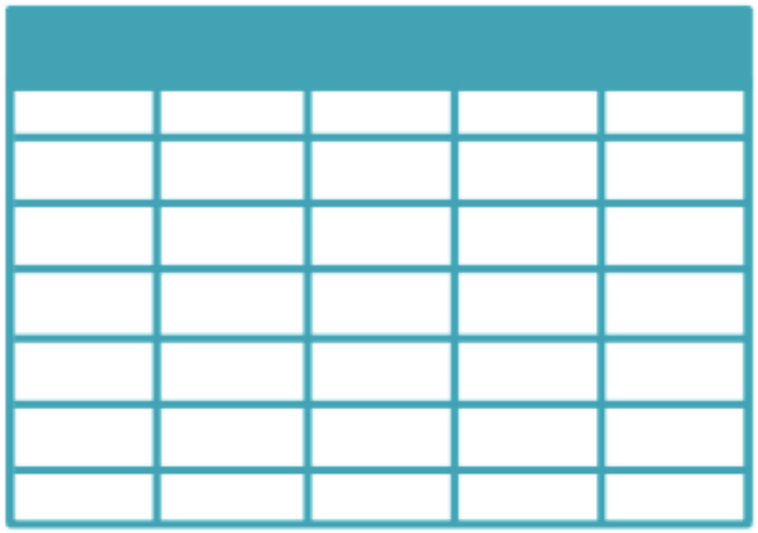


HadoopCon2016

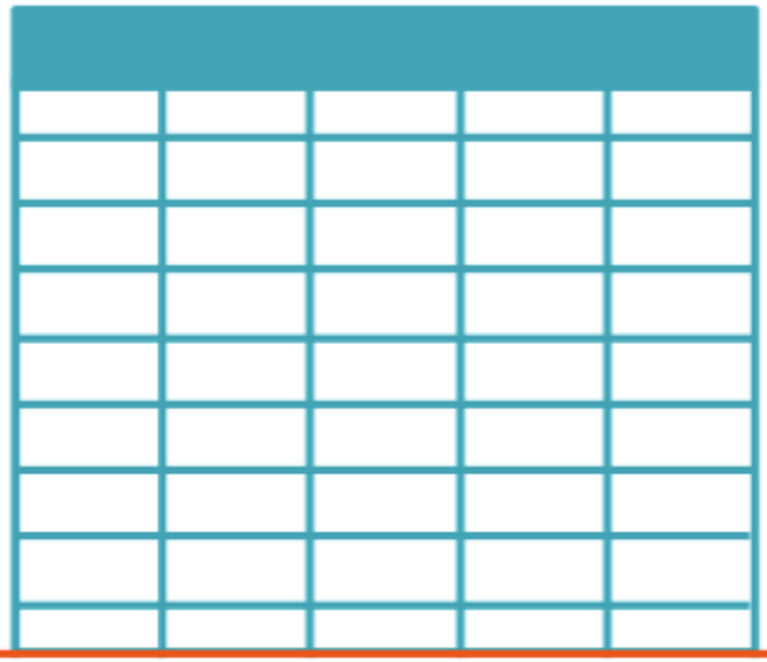
Structured Streaming

DataFrame

Apache Spark 1.3
Static DataFrames

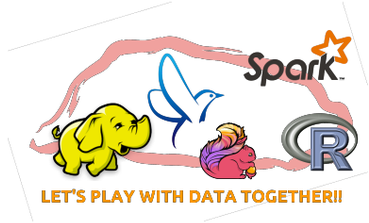


Apache Spark 2.0
Continuous DataFrames



Single API

reference:http://www.slideshare.net/databricks/structuring-spark-dataframes-datasets-and-streaming-62871797



HadoopCon2016

Batch & Continuous

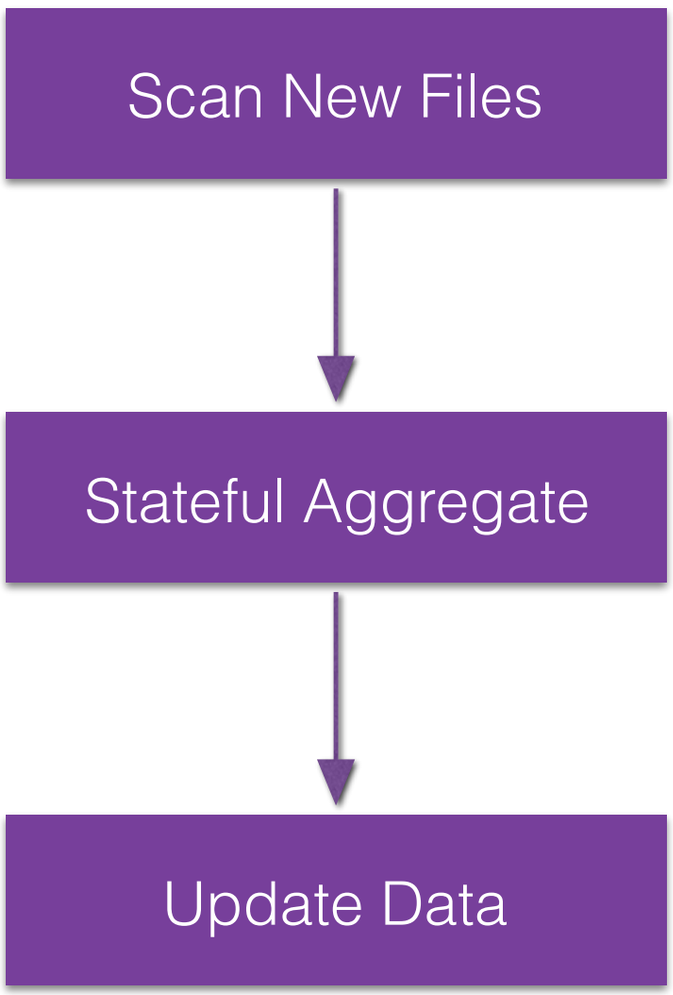
```
logs = spark.read.format(“json").open("s3://logs")

logs.groupBy(logs.user_id).agg(sum(logs.time))
.write.format(“jdbc")
.save("jdbc:mysql//...")
```

Batch Aggregation

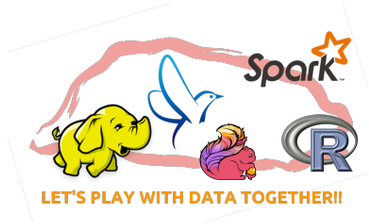
```
logs = spark.read.format("json").stream("s3://logs")

logs.groupBy(logs.user_id).agg(sum(logs.time))
.write.format(“jdbc")
.stream("jdbc:mysql//...")
```

Continuous Aggregation

Continuous Aggregation Flow



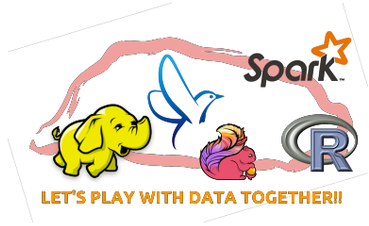
- Cache old file and only read new file.
- Taking the results from all of the past queries and sum them up.
- Continually update data instead of writing once.



Experience with Apache Spark on video games

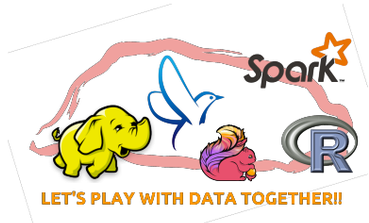

LEAGUE of LEGENDS

reference:**https://lol.garena.tw/game/guide/map/summonersrift**



HadoopCon2016

Is this a presentation about how to win a champion with Spark ?



Unfortunately, no !

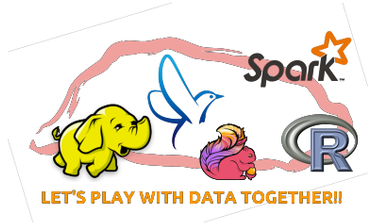
reference:http://www.slideshare.net/SparkSummit/video-games-at-scale-improving-the-gaming-experience-with-apache-spark

Quantity of data

- 67+ million
  monthly active players

- 500+ billion
  data points per day

- 26 petabytes
  data collected since beta



HadoopCon2016

What does Spark do ?

- Spark SQL

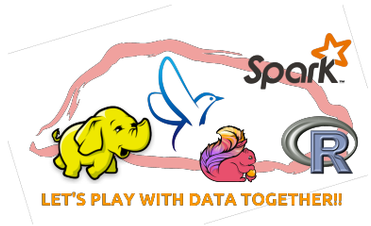
  Data exploration and reporting

- Spark Streaming

  Network performance

- Spark MLlib

  Recommendation system

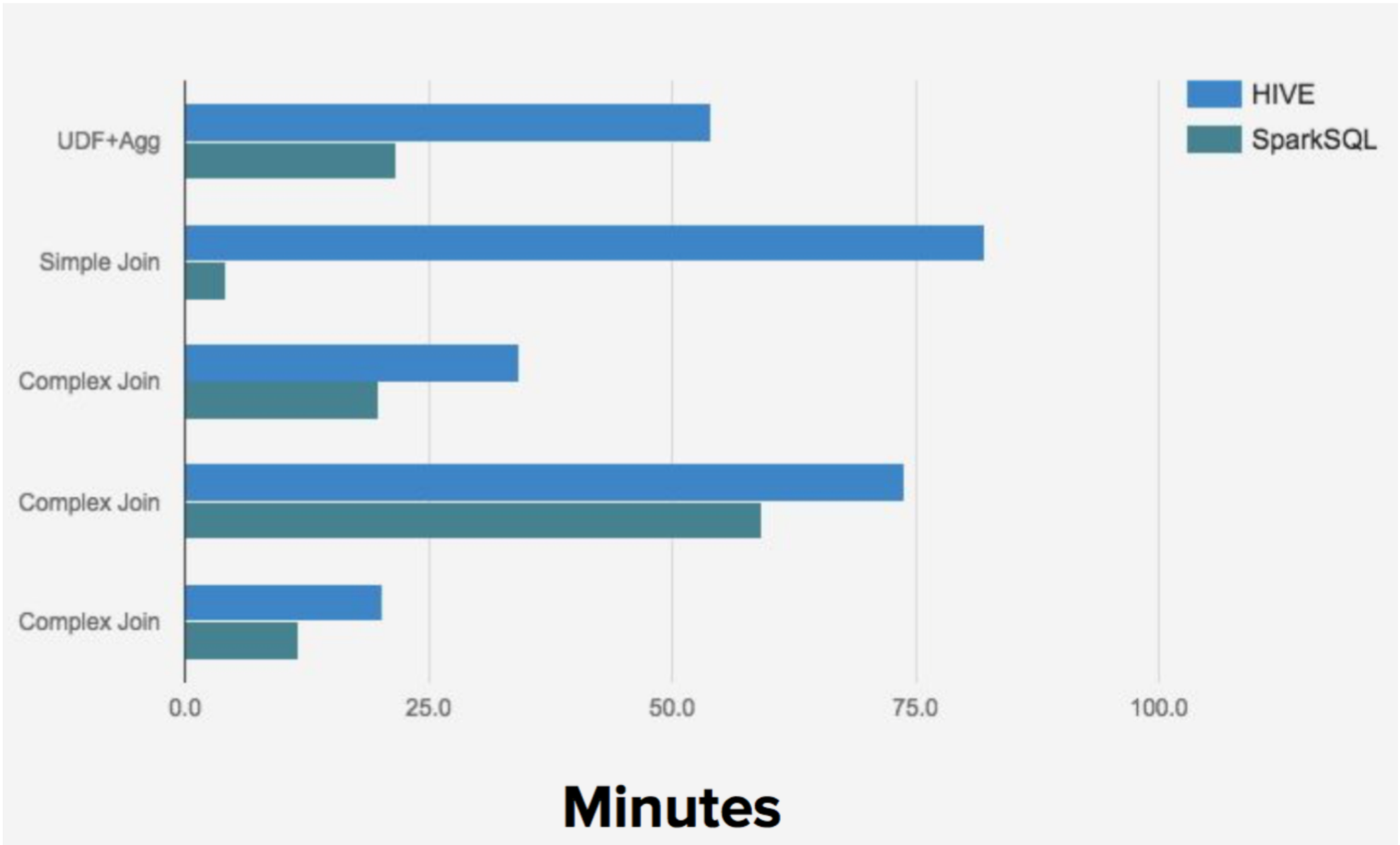


HadoopCon2016

Spark SQL

-Data exploration and reporting

Performance

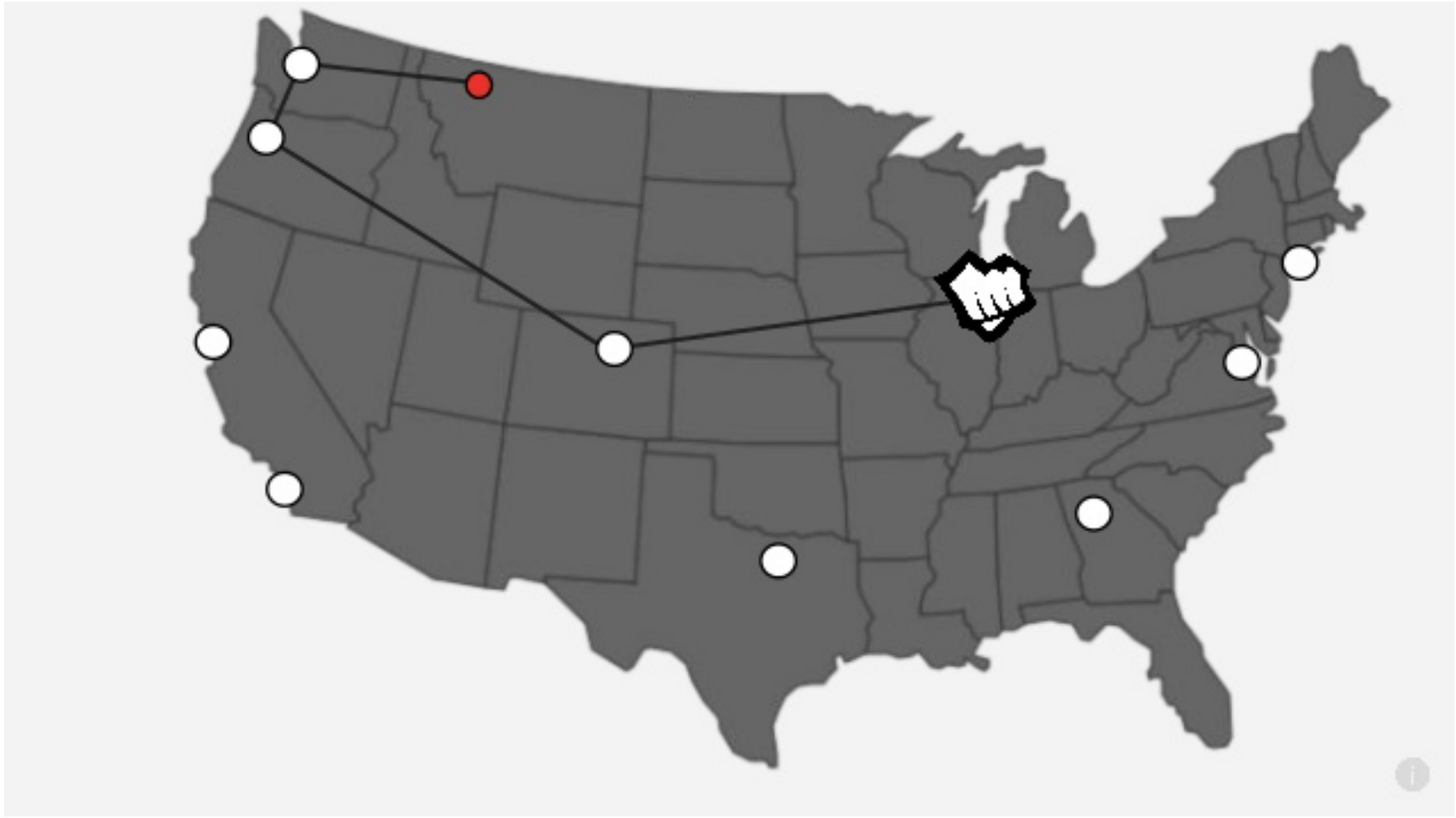


reference:http://www.slideshare.net/SparkSummit/video-games-at-scale-improving-the-gaming-experience-with-apache-spark

Spark Streaming

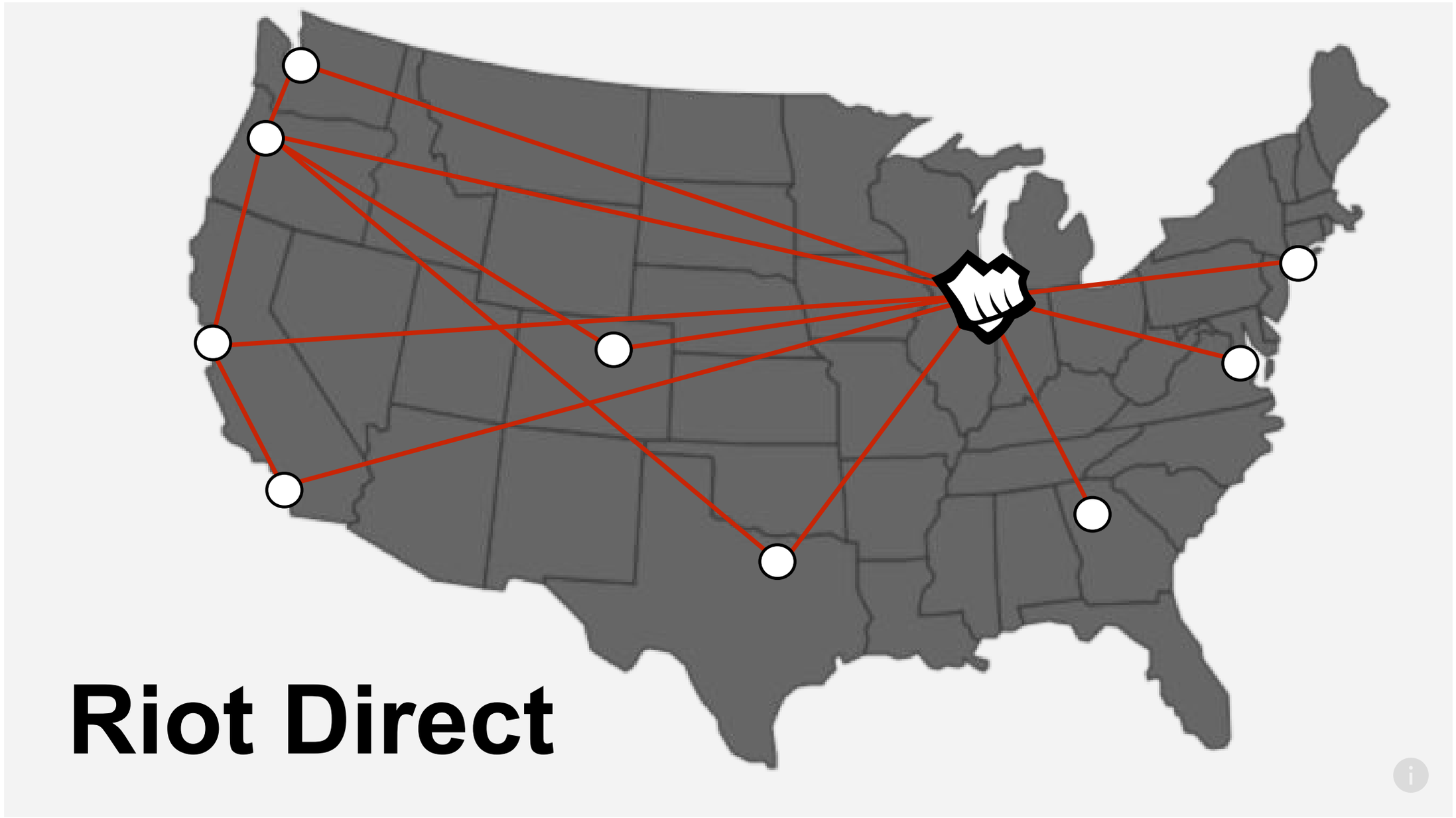
-Network performance

Normal network

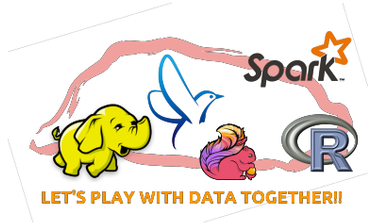


reference:http://www.slideshare.net/SparkSummit/video-games-at-scale-improving-the-gaming-experience-with-apache-spark

Build network



reference:http://www.slideshare.net/SparkSummit/video-games-at-scale-improving-the-gaming-experience-with-apache-spark

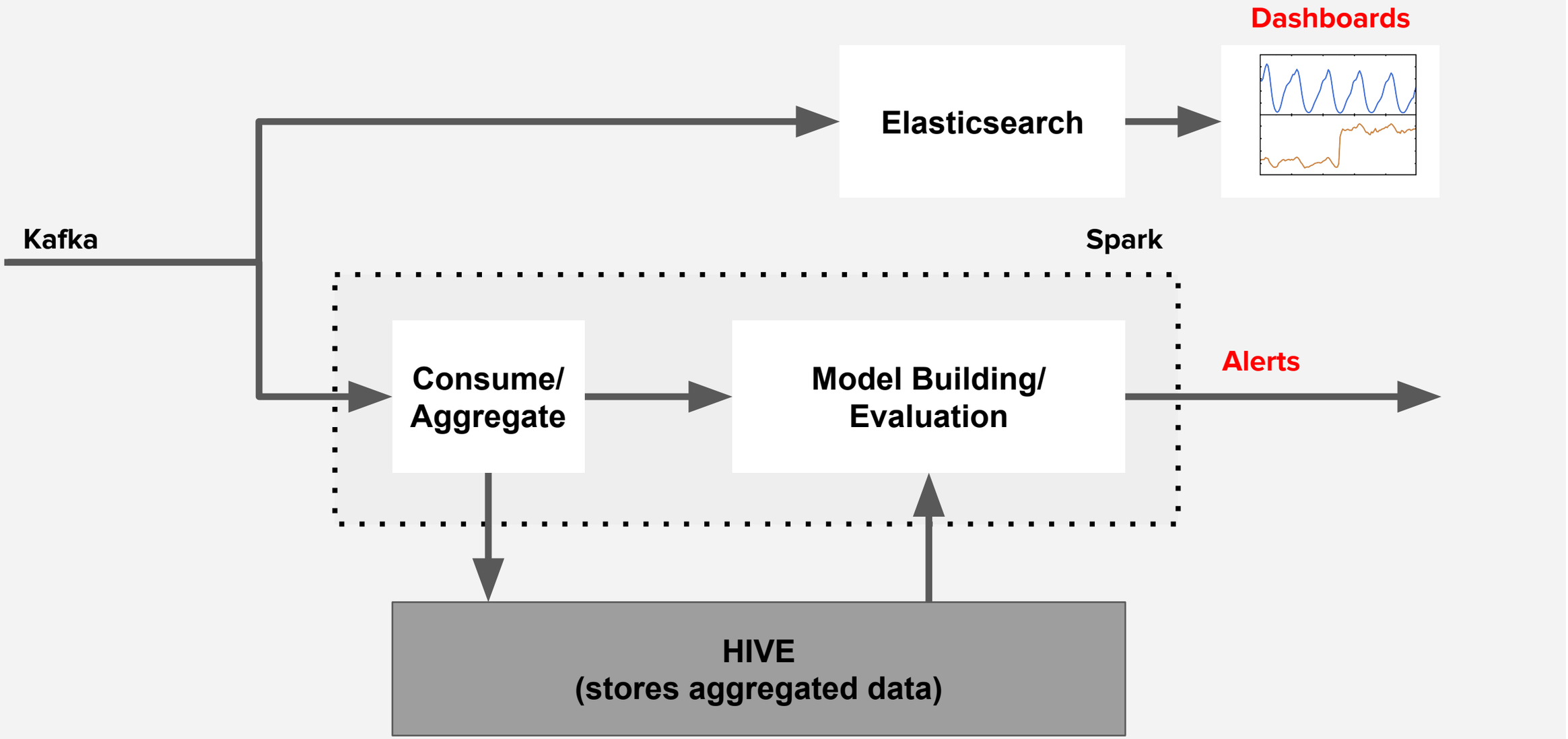


Scope of Data

- 17,000+
  Unique ISPs

- 171,000+
  City/ISP combinations

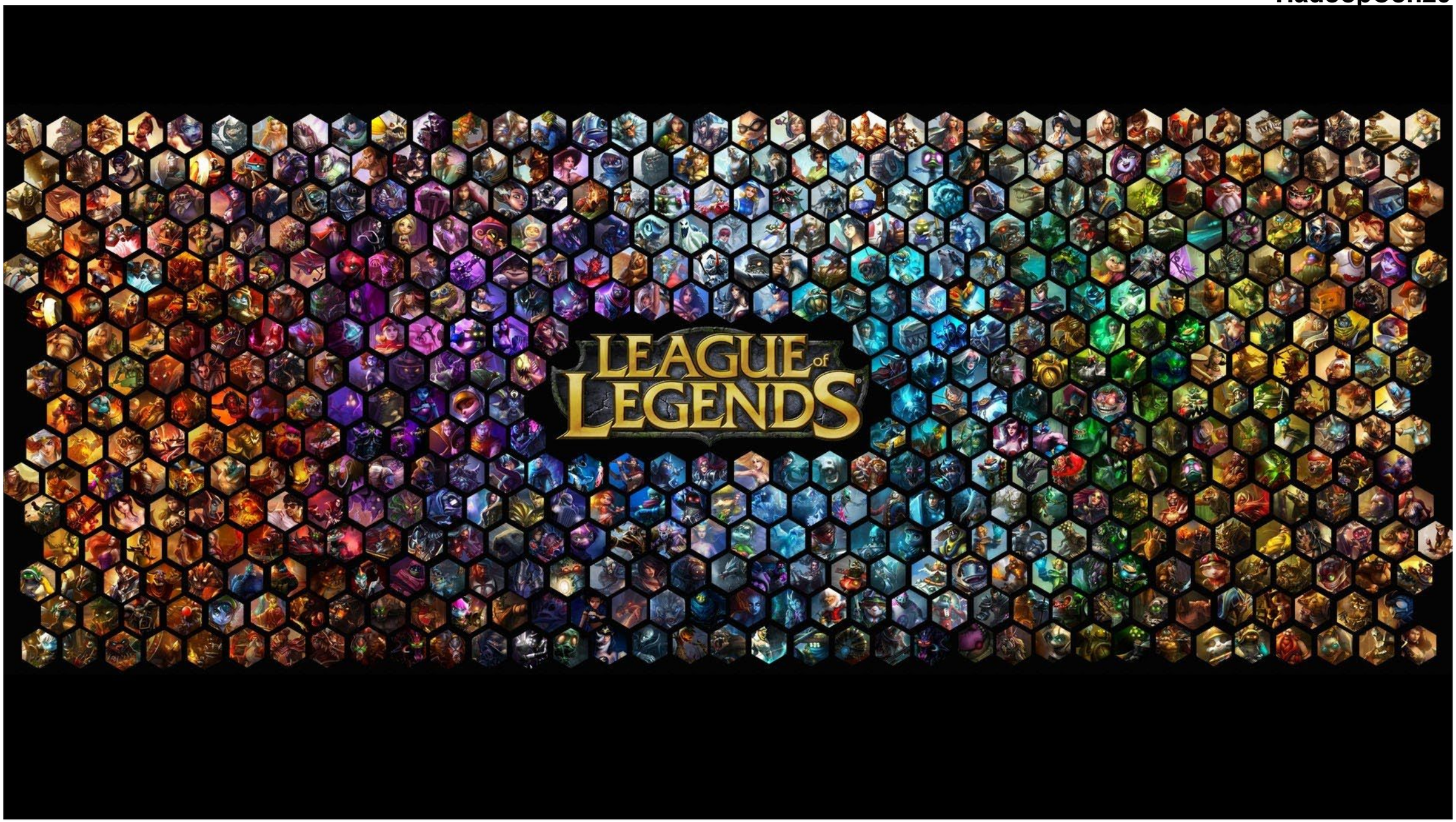
- 250,000+
  Network stat messages per second



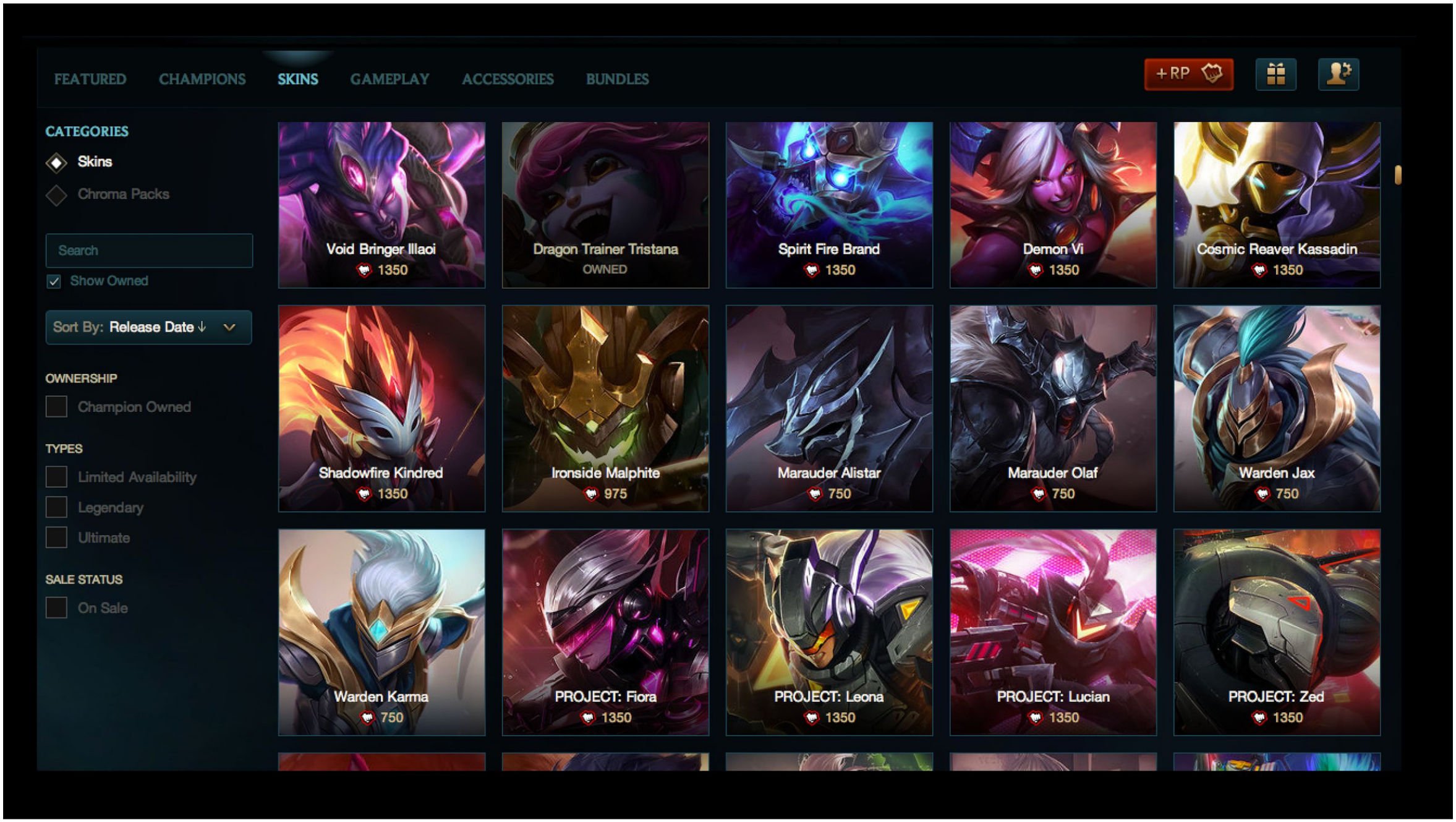
reference:http://www.slideshare.net/SparkSummit/video-games-at-scale-improving-the-gaming-experience-with-apache-spark

Spark MLlib

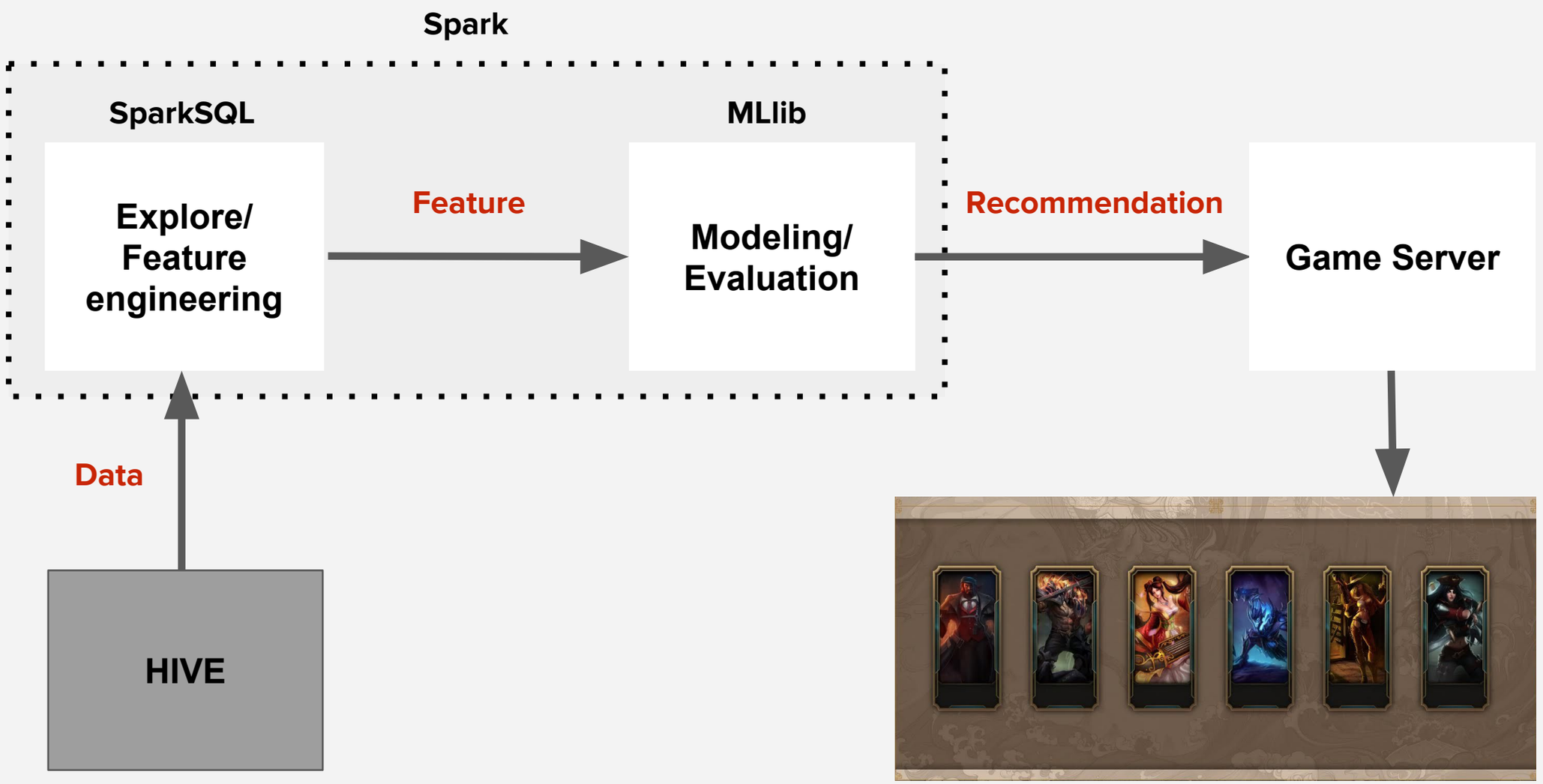
-Recommendation system



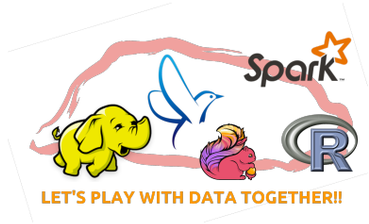
reference:http://www.slideshare.net/SparkSummit/video-games-at-scale-improving-the-gaming-experience-with-apache-spark



reference:http://www.slideshare.net/SparkSummit/video-games-at-scale-improving-the-gaming-experience-with-apache-spark



reference:http://www.slideshare.net/SparkSummit/video-games-at-scale-improving-the-gaming-experience-with-apache-spark



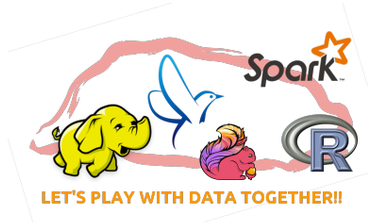
Reference

- Apache Spark
  - http://spark.apache.org/

- Structuring Spark: Dataframes, Datasets And Streaming
  - http://www.slideshare.net/databricks/structuring-spark-dataframes-datasets-and-streaming-62871797

- How to use SparkSession in Apache Spark 2.0
  - https://databricks.com/blog/2016/08/15/how-to-use-sparksession-in-apache-spark-2-0.html

- Video Games at Scale: Improving the gaming experience with Apache Spark
  - http://www.slideshare.net/SparkSummit/video-games-at-scale-improving-the-gaming-experience-with-apache-spark

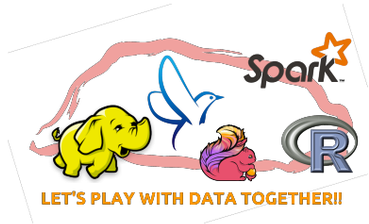


工商時間

- 請加入Spark-Hsinchu Gitter Group
  https://gitter.im/hubertfc/SparkHsinchu

- Gitter app：https://gitter.im/apps

- 與 Spark-Hsinchu Meetup
  https://www.meetup.com/Apache-Spark-Hsinchu/



HadoopCon2016

Thank you

&

Q and A