IoT Applications and Patterns using Apache Spark & Apache Bahir

Luciano Resende

June 14th, 2018

IBM CODE

About me - Luciano Resende

Data Science Platform Architect – IBM – CODAIT

- Have been contributing to open source at ASF for over 10 years
- Currently contributing to : Jupyter Notebook ecosystem, Apache Bahir, Apache Toree, Apache Spark among other projects related to AI/ML platforms

lresende@apache.org

@lresende1975

https://github.com/lresende

https://www.linkedin.com/in/lresende

IBM CODE

Open Source Community Leadership

Apache SOFTWARE FOUNDATION

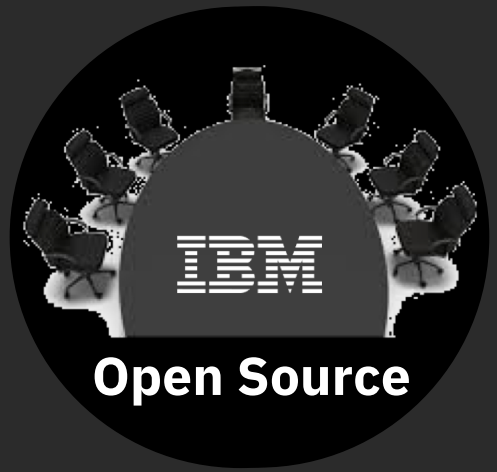
Founding Partner

188+ Project Committers

77+ Projects

amplab UC BERKELEY | riselab UC Berkeley | HYPERLEDGER | R consortium | scalacenter | APACHE SOFTWARE FOUNDATION

Key Open source steering committee memberships



OSS Advisory Board

C O D A I T

IBM CODE

Center for Open Source Data and AI Technologies

CODAIT aims to make AI solutions dramatically easier to create, deploy, and manage in the enterprise

Relaunch of the Spark Technology Center (STC) to reflect expanded mission

IBM CODE

CODAIT

codait.org

codait (French)

= coder/coded

https://m.interglot.com/fr/en/codait

Agenda

Introductions

- Apache Spark
- Apache Bahir

IoT Applications

Live Demo

Summary

References

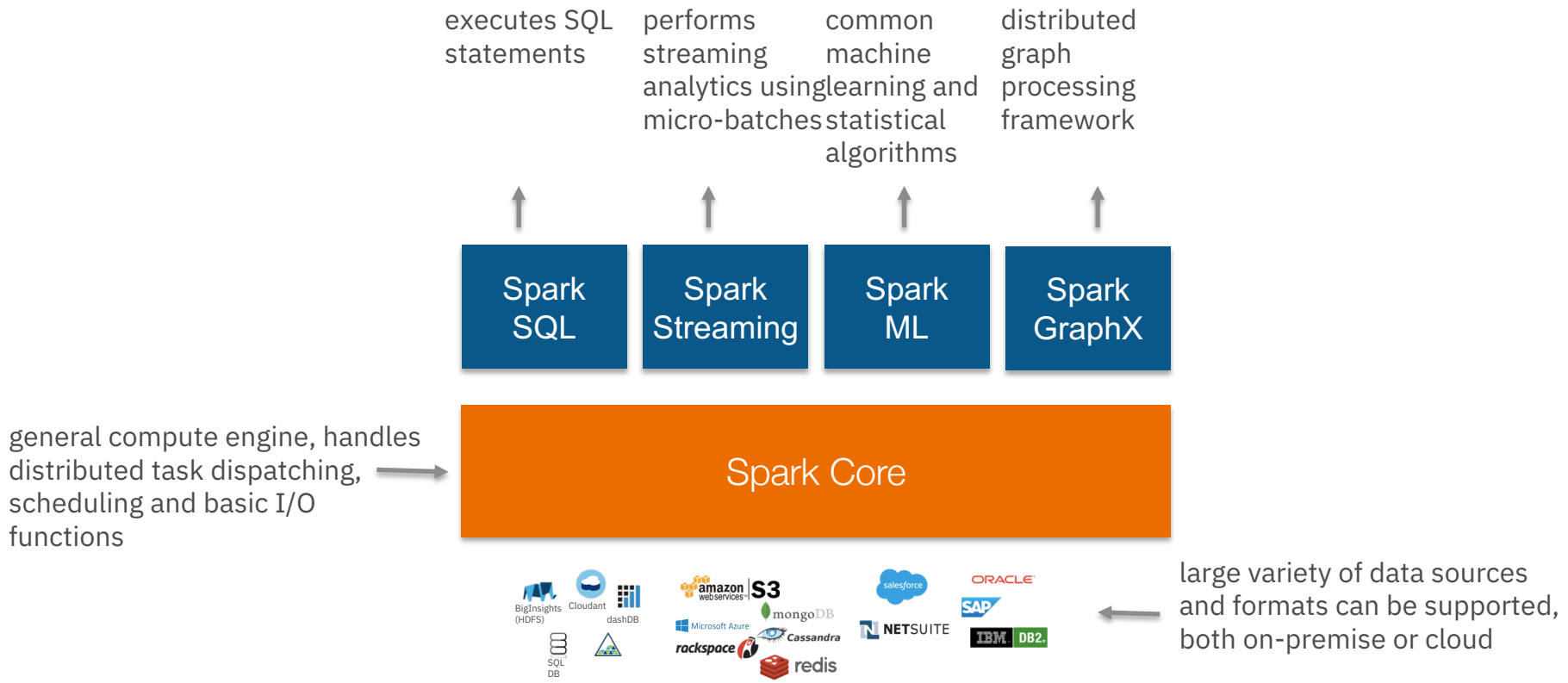
Q&A

IBM CODE

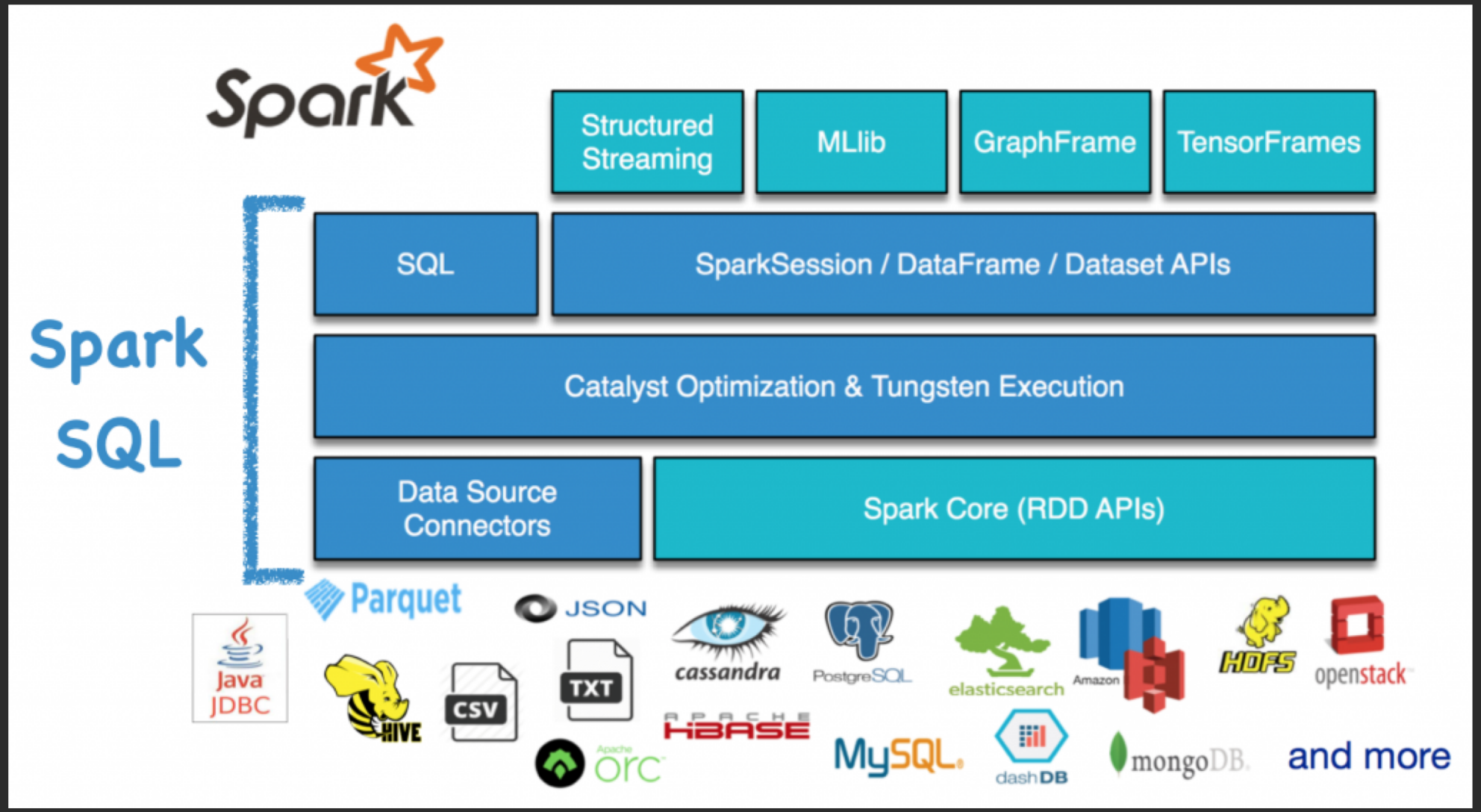
Apache Spark

IBM
CODE

Apache Spark Introduction

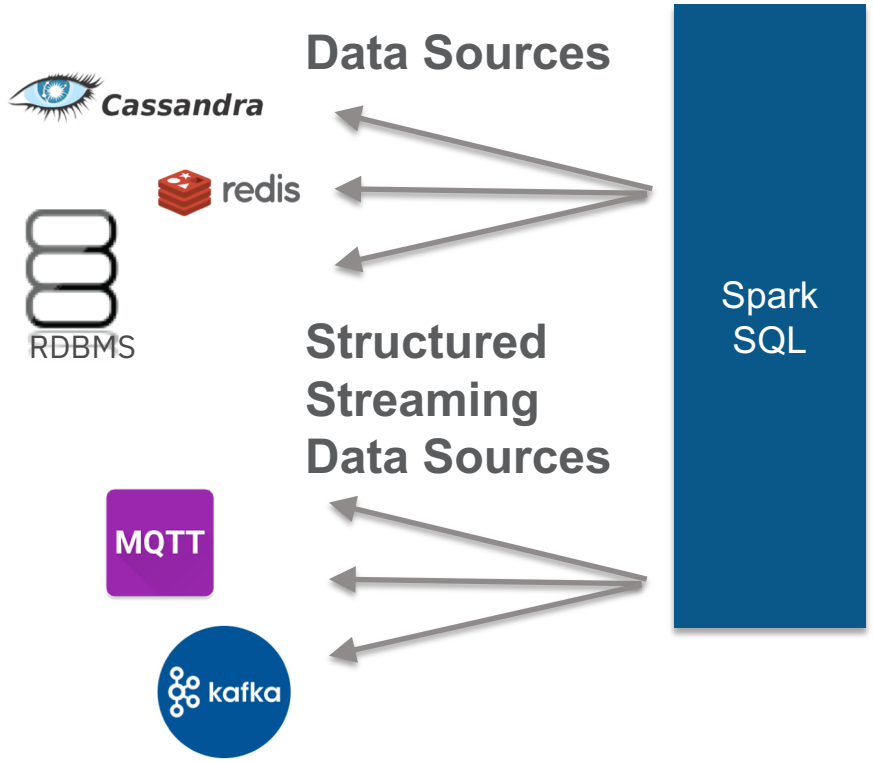


Apache Spark Evolution



IBM CODE

Apache Spark – Spark SQL



Unified data access APIS: Query structured data sets with SQL or Dataset/DataFrame APIs

Fast, familiar query language across all of your enterprise data

Apache Spark – Spark SQL

You can run SQL statement with SparkSession.sql(...) interface:

```
val spark = SparkSession.builder()
  .appName(“Demo”)
  .getOrCreate()

spark.sql(“create table T1 (c1 int, c2 int) stored as parquet”)

val ds = spark.sql(“select * from T1”)
```

You can further transform the resultant dataset:

```
val ds1 = ds.groupBy(“c1”).agg(“c2”-> “sum”)

val ds2 = ds.orderBy(“c1”)
```

The result is a DataFrame / Dataset[Row]

ds.show() displays the rows

IBM CODE

Apache Spark – Spark SQL

You can read from data sources using SparkSession.read.format(...)

```
val spark = SparkSession.builder()
  .appName(“Demo”)
  .getOrCreate()

case class Bank(age: Integer, job: String, marital: String, education: String,  balance: Integer)

// loading csv data to a Dataset of Bank type

val bankFromCSV = spark.read.csv(“hdfs://localhost:9000/data/bank.csv").as[Bank]

// loading JSON data to a Dataset of Bank type

val bankFromJSON = spark.read.json(“hdfs://localhost:9000/data/bank.json").as[Bank]

// select a column value from the Dataset
```

bankFromCSV.select(‘age).show() will return all rows of column “age” from this dataset.

IBM CODE

Apache Spark – Spark SQL

You can also configure a specific data source with specific options

```
val spark = SparkSession.builder()
  .appName(“Demo”)
  .getOrCreate()

case class Bank(age: Integer, job: String, marital: String, education: String,  balance: Integer)

// loading csv data to a Dataset of Bank type

val bankFromCSV = sparkSession.read

  .option("header", ”true") // Use first line of all files as header

  .option("inferSchema", ”true") // Automatically infer data types

  .option("delimiter", " ")

  .csv("/users/lresende/data.csv”)

  .as[Bank]

bankFromCSV.select(‘age).show()  // will return all rows of column “age” from this dataset.
```

IBM CODE

Apache Spark – Spark SQL – Data Sources

Data Sources under the covers

- Data source registration (e.g. spark.read.datasource)
- Provide BaseRelation implementation
  - That implements support for table scans:
    - TableScans, PrunedScan, PrunedFilteredScan, CatalystScan
- Detailed information available at
  - https://developer.ibm.com/code/2016/11/10/exploring-apache-spark-datasource-api/

IBM CODE

Apache Spark – Spark SQL – Data Sources

Data Sources V1 Limitations

- Leak upper-level API in the data source (DataFrame/SQLContext)
- Hard to extend the Data Sources API for more optimizations
- Zero transaction guarantee in the write APIs
- Limited Extensibility

Apache Spark – Spark SQL – Data Sources

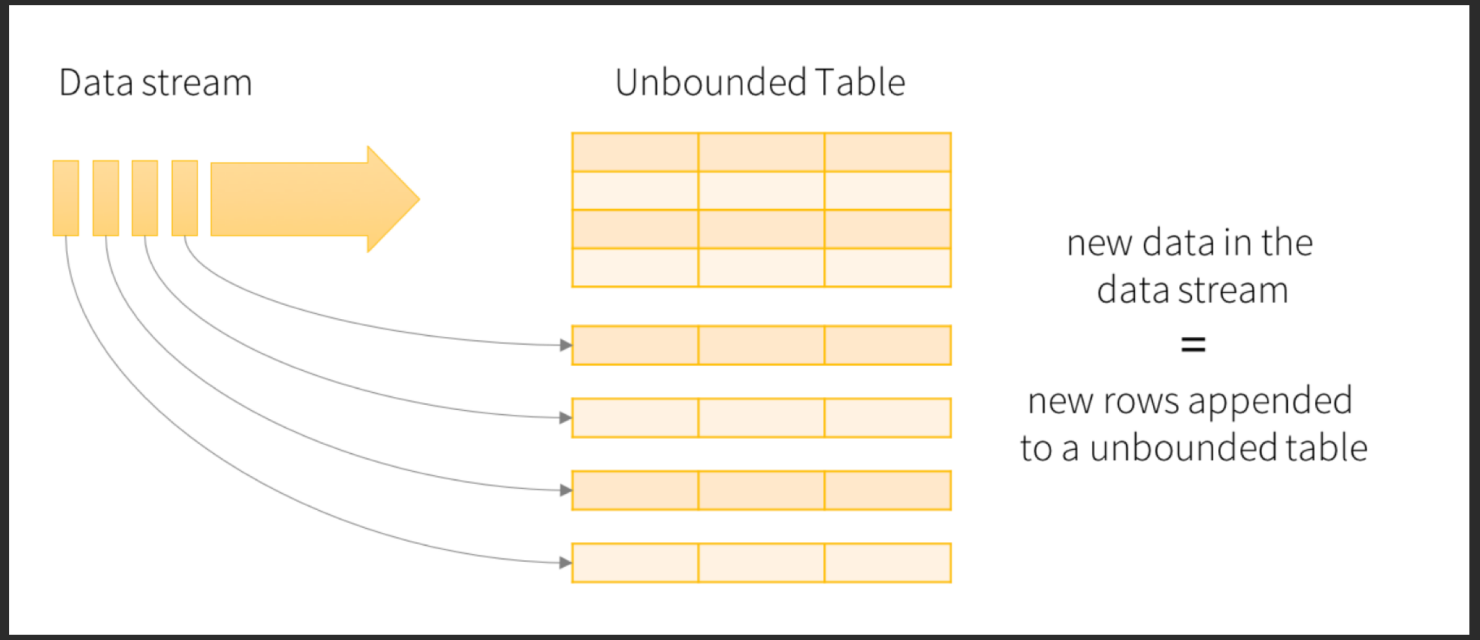
Data Sources V2

- Support for row-based scan and columnar scan
- Column pruning and filter push-down
- Can report basic statistics and data partitioning
- Transactional write API
- Streaming source and sink support for micro-batch and continuous mode
- Detailed information available at
  - https://developer.ibm.com/code/2018/04/16/introducing-apache-spark-data-sources-api-v2/

IBM CODE

Apache Spark – Spark SQL Structured Streaming

Unified programming model for streaming, interactive and batch queries



Considers the data stream as unbounded table

IBM CODE

Image source: https://spark.apache.org/docs/latest/structured-streaming-programming-guide.html

Apache Spark – Spark SQL Structured Streaming

SQL regular APIs

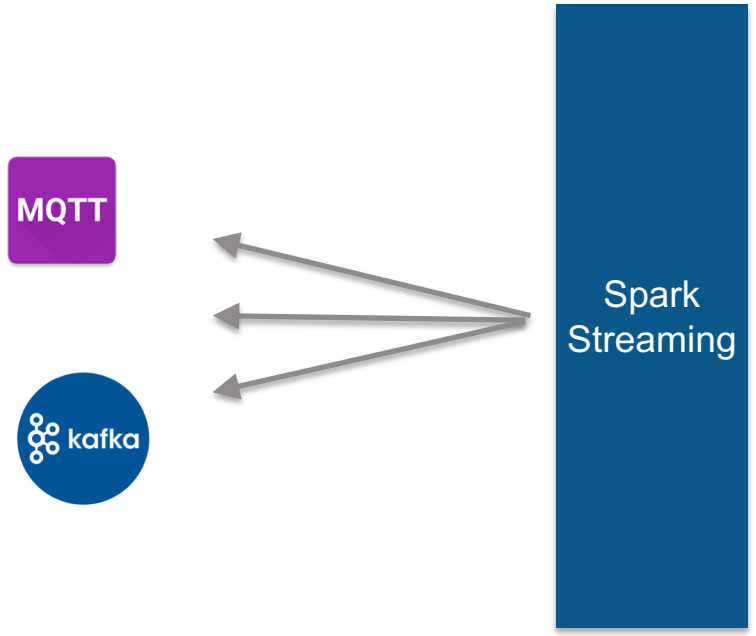
```
val spark = SparkSession.builder()
  .appName(“Demo”)
  .getOrCreate()

 val input = spark.read
     .schema(schema)
     .format(”csv")
     .load(”input-path")

val result = input
     .select(”age”)
     .where(”age > 18”)

result.write
     .format(”json”)
     . save(” dest-path”)
```

Structured Streaming APIs

```
val spark = SparkSession.builder()
  .appName(“Demo”)
  .getOrCreate()
 val input = spark.readStream
     .schema(schema)
     .format(”csv")
     .load(”input-path")
val result = input
     .select(”age”)
     .where(”age > 18”)
result.write
     .format(”json”)
     . startStream(” dest-path”)
```

IBM CODE

Apache Spark – Spark Streaming



Micro-batch event processing for near-real time analytics

e.g. Internet of Things (IoT) devices, Twitter feeds, Kafka (event hub), etc.

No multi-threading or parallel process programming required

IBM CODE

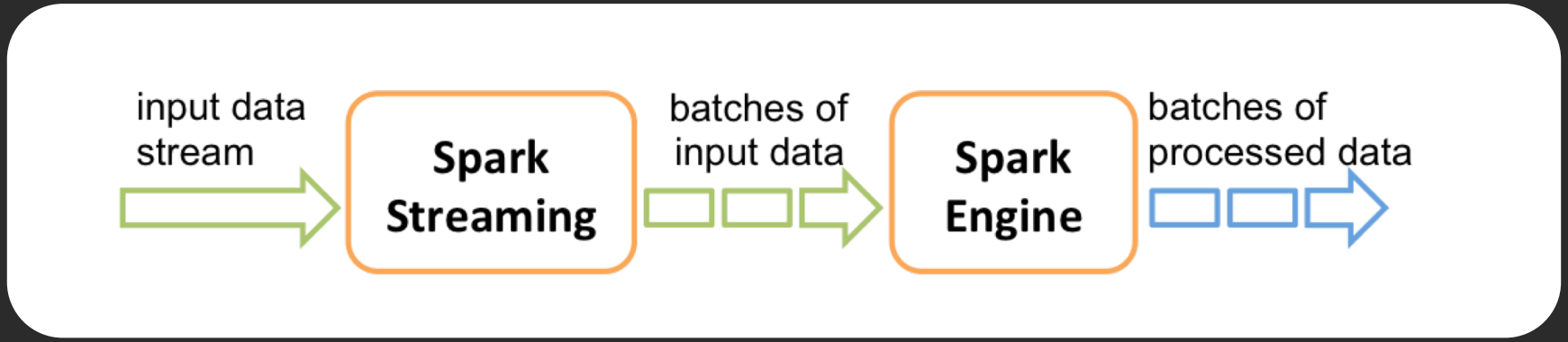
Apache Spark – Spark Streaming

Also known as discretized stream or DStream

Abstracts a continuous stream of data

Based on micro-batching

Based on RDDs



IBM CODE

Apache Spark – Spark Streaming

```
 val sparkConf = new SparkConf()
   .setAppName("MQTTWordCount")


val ssc = new StreamingContext(sparkConf, Seconds(2))

val lines = MQTTUtils.createStream(ssc, brokerUrl, topic, StorageLevel.MEMORY_ONLY_SER_2)


val words = lines.flatMap(x => x.split(" "))

val wordCounts = words.map(x => (x, 1)).reduceByKey(_ + _)

wordCounts.print()


ssc.start()

ssc.awaitTermination()
```

IBM CODE

Apache Bahir

Origins of the Apache Bahir Project

MAY/2016: Established as a top-level Apache Project.

- PMC formed by Apache Spark committers/pmc, Apache Members
- Initial contributions imported from Apache Spark

AUG/2016: Apache Flink community join Apache Bahir

- Initial contributions of Flink extensions
- In October 2016 Robert Metzger elected committer

Origins of the Bahir name

Naming an Apache Project is a science !!!

- We needed a name that wasn’t used yet
- Needed to be related to Spark

We ended up with : Bahir

- A name of Arabian origin that means Sparkling,
- Also associated with a guy who succeeds at everything

Why Apache Bahir

It’s an Apache project

- And if you are here, you know what it means

Benefits of curating your extensions at Apache Bahir

- Apache Governance
- Apache License
- Apache Community
- Apache Brand

IBM CODE

Why Apache Bahir

Flexibility

- Release flexibility
  - Bounded to platform or component release

Shared infrastructure

- Release, CI, etc

Shared knowledge

- Collaborate with experts on both platform and component areas

IBM CODE

Bahir extensions for Apache Spark

MQTT – Enables reading data from MQTT Servers using Spark Streaming or Structured streaming.

- http://bahir.apache.org/docs/spark/current/spark-sql-streaming-mqtt/
- http://bahir.apache.org/docs/spark/current/spark-streaming-mqtt/

Couch DB/Cloudant – Enables reading data from CouchDB/Cloudant using Spark SQL and Spark Streaming.

Twitter – Enables reading social data from twitter using Spark Streaming.

- http://bahir.apache.org/docs/spark/current/spark-streaming-twitter/

Akka – Enables reading data from Akka Actors using Spark Streaming or Structured Streaming.

- http://bahir.apache.org/docs/spark/current/spark-streaming-akka/

ZeroMQ – Enables reading data from ZeroMQ using Spark Streaming.

- http://bahir.apache.org/docs/spark/current/spark-streaming-zeromq/

IBM CODE

Bahir extensions for Apache Spark

Google Cloud Pub/Sub – Add spark streaming connector to Google Cloud Pub/Sub

Apache Spark extensions in Bahir

Adding Bahir extensions into your application

- Using SBT

```
libraryDependencies += "org.apache.bahir" %% "spark-streaming-mqtt" % "2.2.0”
```

- Using Maven

```
<dependency>
  <groupId>org.apache.bahir</groupId>
  <artifactId>spark-streaming-mqtt_2.11 </artifactId>
  <version>2.2.0</version>
</dependency>
```

IBM CODE

Apache Spark extensions in Bahir

Submitting applications with Bahir extensions to Spark

- Spark-shell

  bin/spark-shell --packages org.apache.bahir:spark-streaming_mqtt_2.11:2.2.0 .....

- Spark-submit

  bin/spark-submit --packages org.apache.bahir:spark-streaming_mqtt_2.11:2.2.0 .....

IBM CODE

Internet of Things - IoT

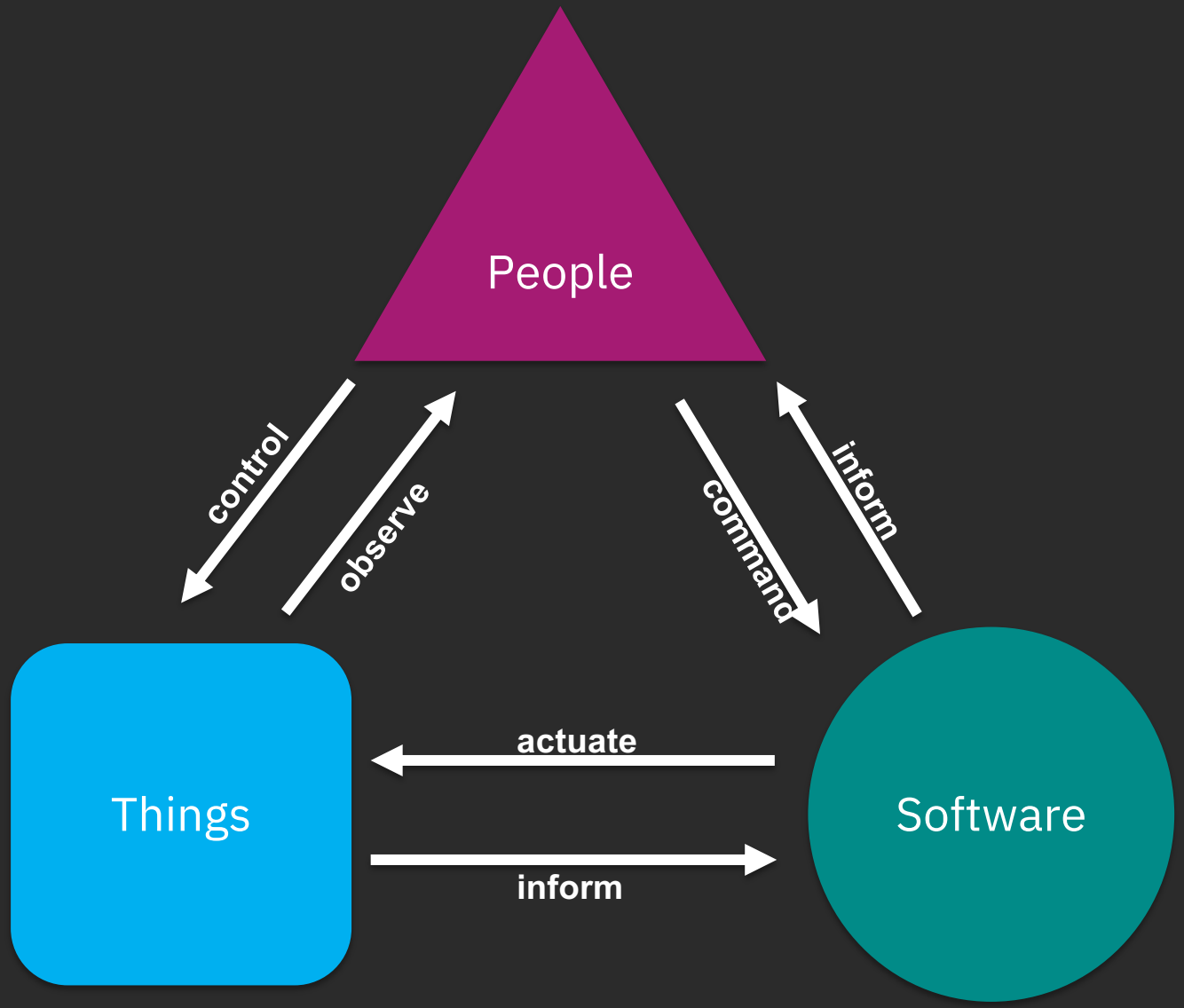
IBM CODE

IoT – Definition by Wikipedia

The Internet of things (IoT) is the network of physical devices, vehicles, home appliances, and other items embedded with electronics, software sensors, actuators, and network connectivity which enable these objects to connect and exchange data.

IBM CODE

IoT – Interaction between multiple entities



IoT Patterns – Some of them ...

- Remote control
- Security analysis
- Edge analytics
- Historical data analysis
- Distributed Platforms
- Real-time decisions

MQTT – M2M / IoT Connectivity Protocol

Connect
+
Publish
+
Subscribe

Open spec

+ 40 client implementations

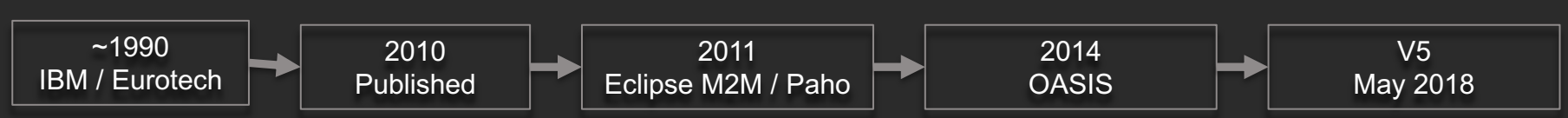
Minimal overhead

Header
2-4 bytes (publish)
14 bytes (connect)

Tiny Clients (Java 170KB)

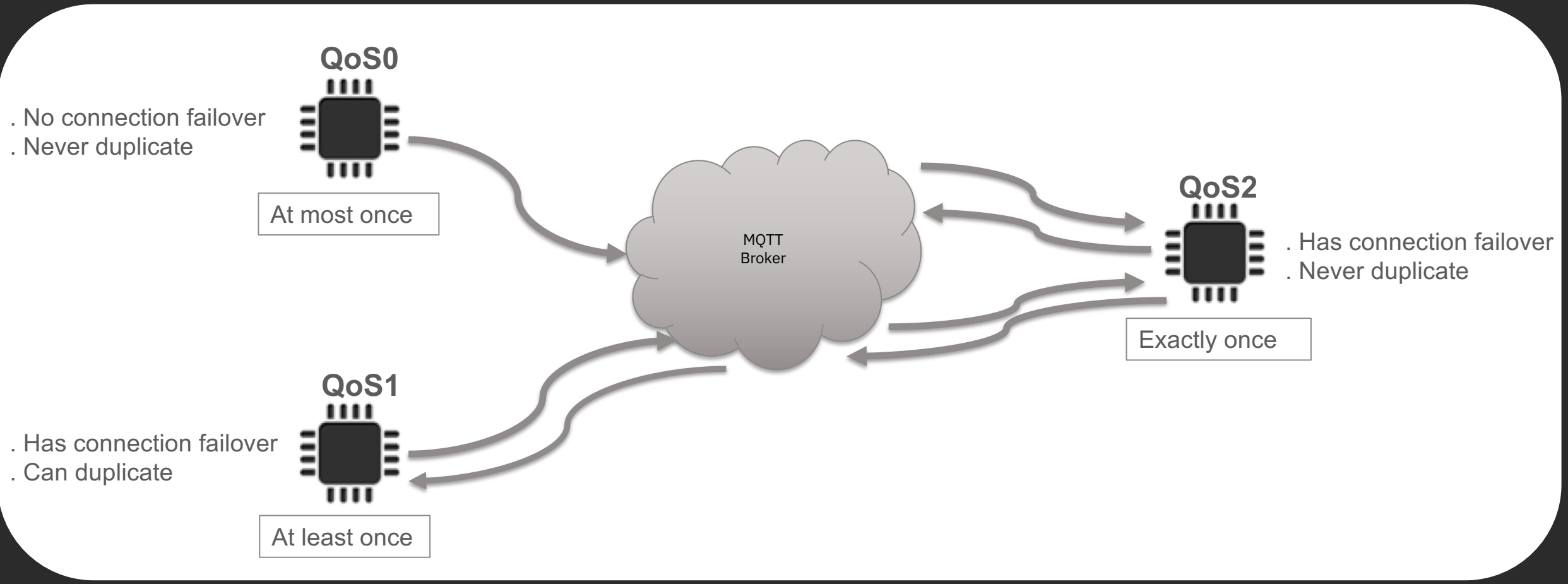
MQTT

History



IBM CODE

MQTT – Quality of Service



MQTT – World usage

Smart Home Automation

Messaging

Notable Mentions:

- IBM IoT Platform
- AWS IoT
- Microsoft IoT Hub
- Facebook Messanger

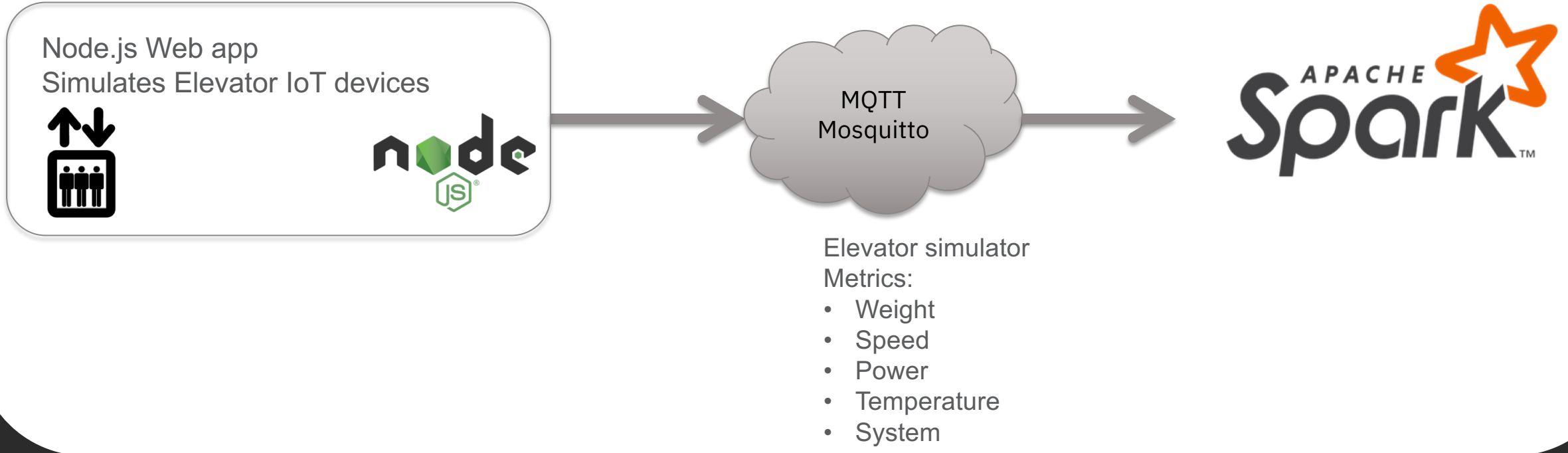
IBM CODE

Live Demo

IBM
CODE

IoT Simulator using MQTT

The demo environment
https://github.com/lresende/bahir-iot-demo



Summary

IBM CODE

Summary – Take away points

Apache Spark

- IoT Analytics Runtime with support for ”Continuous Applications”

Apache Bahir

- Bring access to IoT data via supported connectors (e.g. MQTT)

IoT Applications

- Using Spark and Bahir to start processing IoT data in near real time using Spark Streaming and Spark Structured Streaming

IBM CODE

Join the Apache Bahir community

References

Apache Bahir

http://bahir.apache.org

Documentation for Apache Spark extensions

http://bahir.apache.org/docs/spark/current/documentation/

Source Repositories

https://github.com/apache/bahir

https://github.com/apache/bahir-website

Demo Repository

https://github.com/lresende/bahir-iot-demo

Image source: http://az616578.vo.msecnd.net/files/2016/03/21/635941249931013850155786752 9_thank-you-1400x800-c-default.gif