Livy: A REST Web Service for Spark

Pravin Mittal, Microsoft

Anand Iyer, Cloudera

cloudera

Microsoft

Reduce friction to use Spark

while

maintaining all its power and flexibility

cloudera

Microsoft

What is Livy?

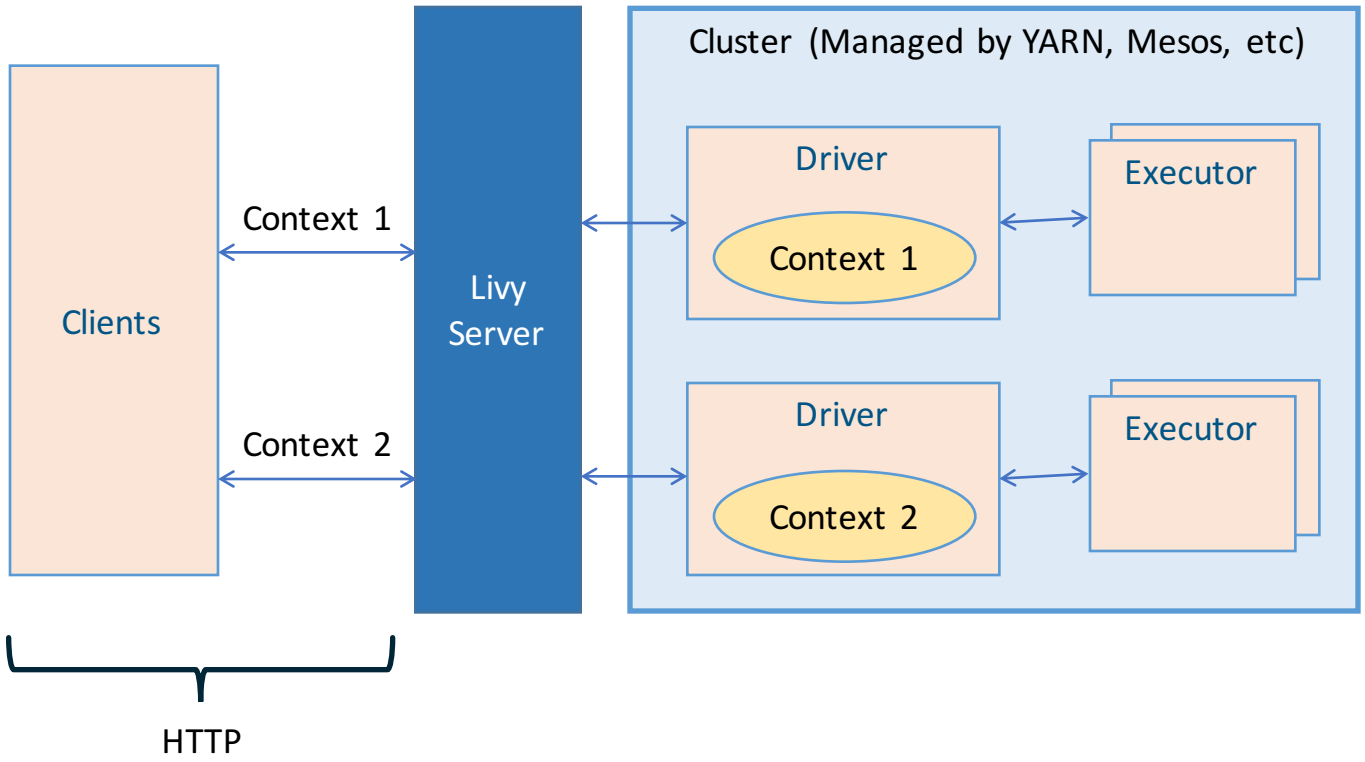
A Service that manages long running Spark Contexts in your cluster

- Open Source Apache Licensed
- REST based interface
- Lets you manage multiple Spark Contexts
- Fine grained job submission
- Retrieve job results over REST asynchronously or synchronously
- Client APIs in java, scala and soon in python

cloudera

Microsoft

What is Livy?



cloudera

Microsoft

Spark on Azure HDInsight

Fully Managed Service

- 100% open source Apache Spark and Hadoop bits
- Latest releases of Spark
- Fully supported by Microsoft and Hortonworks
- 99.9% Azure Cloud SLA; 24/7 Managed Service
- Certifications: PCI, ISO 27018, SOC, HIPAA, EU-MC

Optimized for experimentation and development

- Jupyter Notebooks (scala, python, automatic data visualizations)
- IntelliJ plugin (job submission, remote debugging)
- ODBC connector for Power BI, Tableau, Qlik, SAP, Excel, etc

cloudera

Microsoft

Make Spark Simple - Integrated with Azure Ecosystem

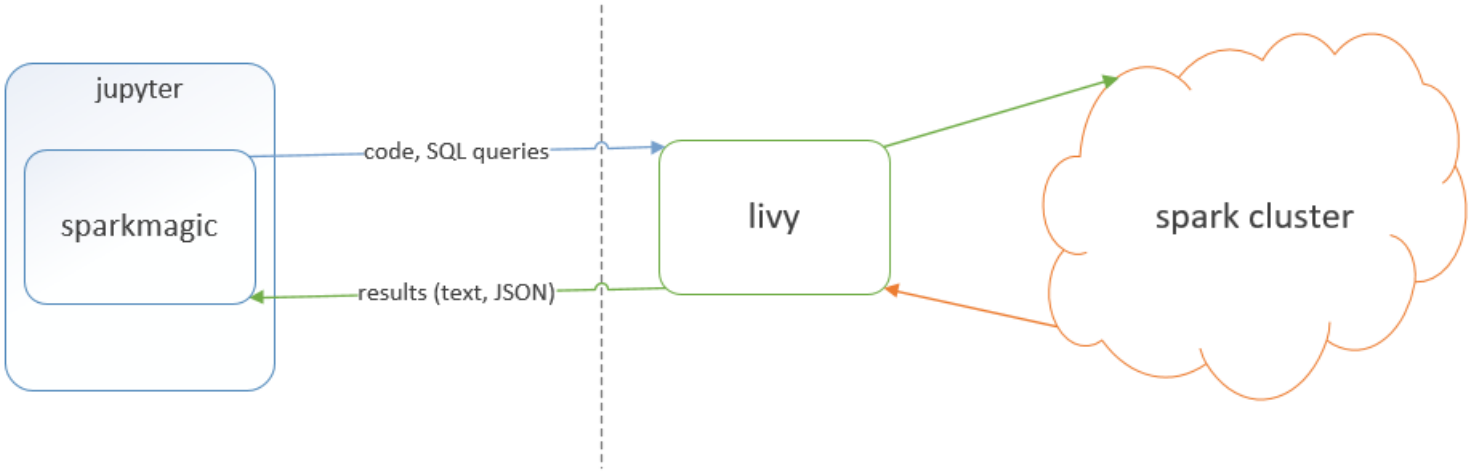
- Microsoft R Server - Multi-threaded math libraries and transparent parallelization in R Server means handling up to 1000x more data and up to 50x faster speeds than open source R. This is based on open source R, it does require any change to R scripts
- Azure Data Lake Store – HDFS for the cloud, optimized for massive throughput, Ultra-high capacity, Low Latency, Secure ACL support
- Azure Data Factory orchestrates Spark ETL pipeline
- PowerBI connector for Spark for rich visualization. New in Power BI is a streaming connector allowing you to publish real-time events from Spark Streaming directly to Power BI.
- EventsHub connector as a data source for Spark streaming
- Azure SQL Datawarehouse & Hbase connector for fast & scalable storage

cloudera

Microsoft

Jupyter-Spark Integration via Livy

- Sparkmagic is an open source library that Microsoft is incubating under the Jupyter Incubator program
- Thousands of Spark clusters in production providing feedback to further improve the experience



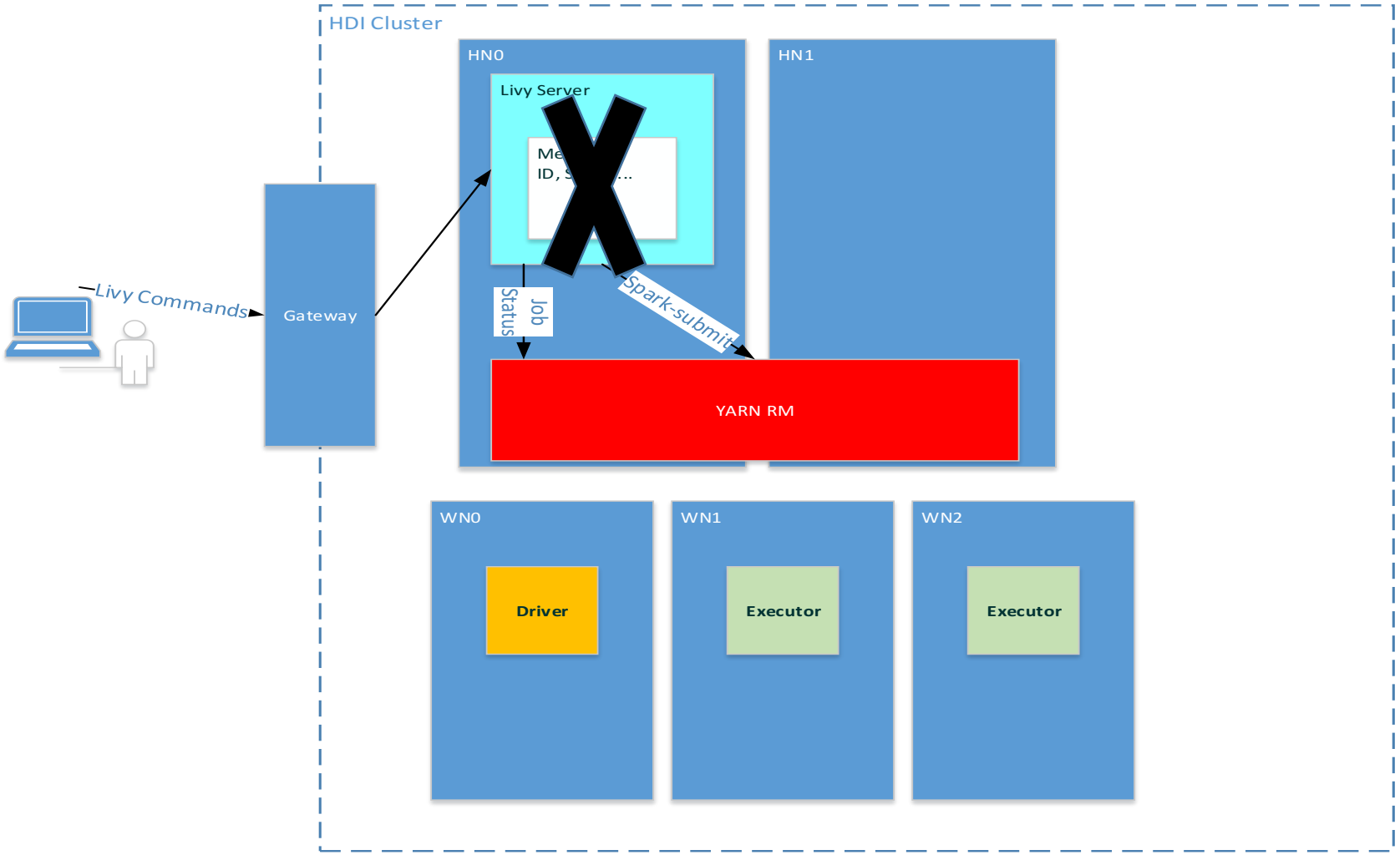
https://github.com/jupyter-incubator/sparkmagic

cloudera

Microsoft

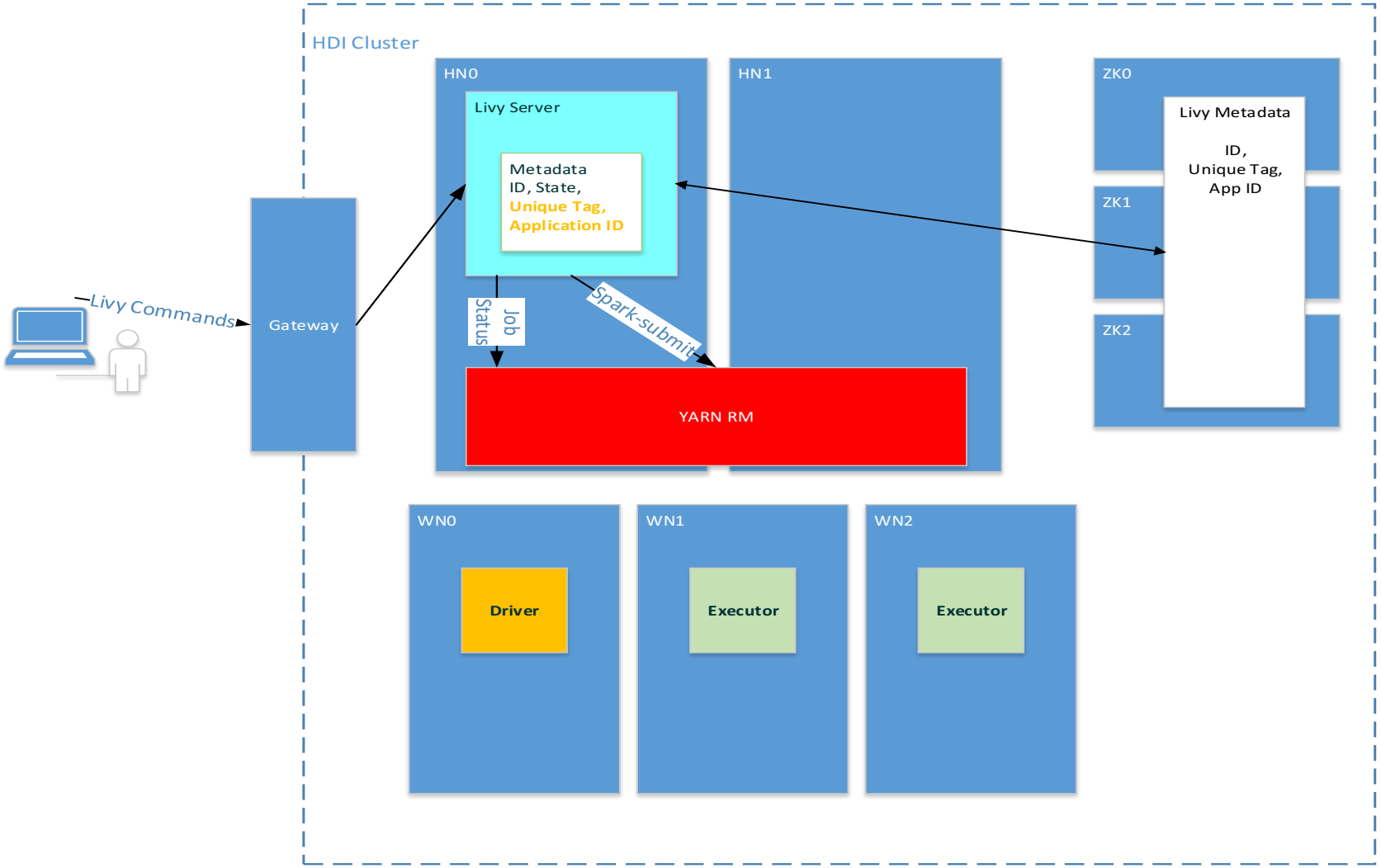
Architectural Advantages of Jupyter integration via Livy

- Run Spark code completely remotely; no Spark components need to be installed on the Jupyter server
- Multi-language support; the Python, Scala and R kernels are equally feature-rich
- Support for multiple endpoints; you can use a single notebook to start multiple Spark jobs in different languages and against different remote clusters
- Easy integration with any Python library for data science or visualization, like Pandas or Plotly



cloudera

Microsoft



cloudera

Microsoft

DEMO

cloudera

Microsoft

Resources

https://github.com/aggFTW/sparksummit2016

cloudera

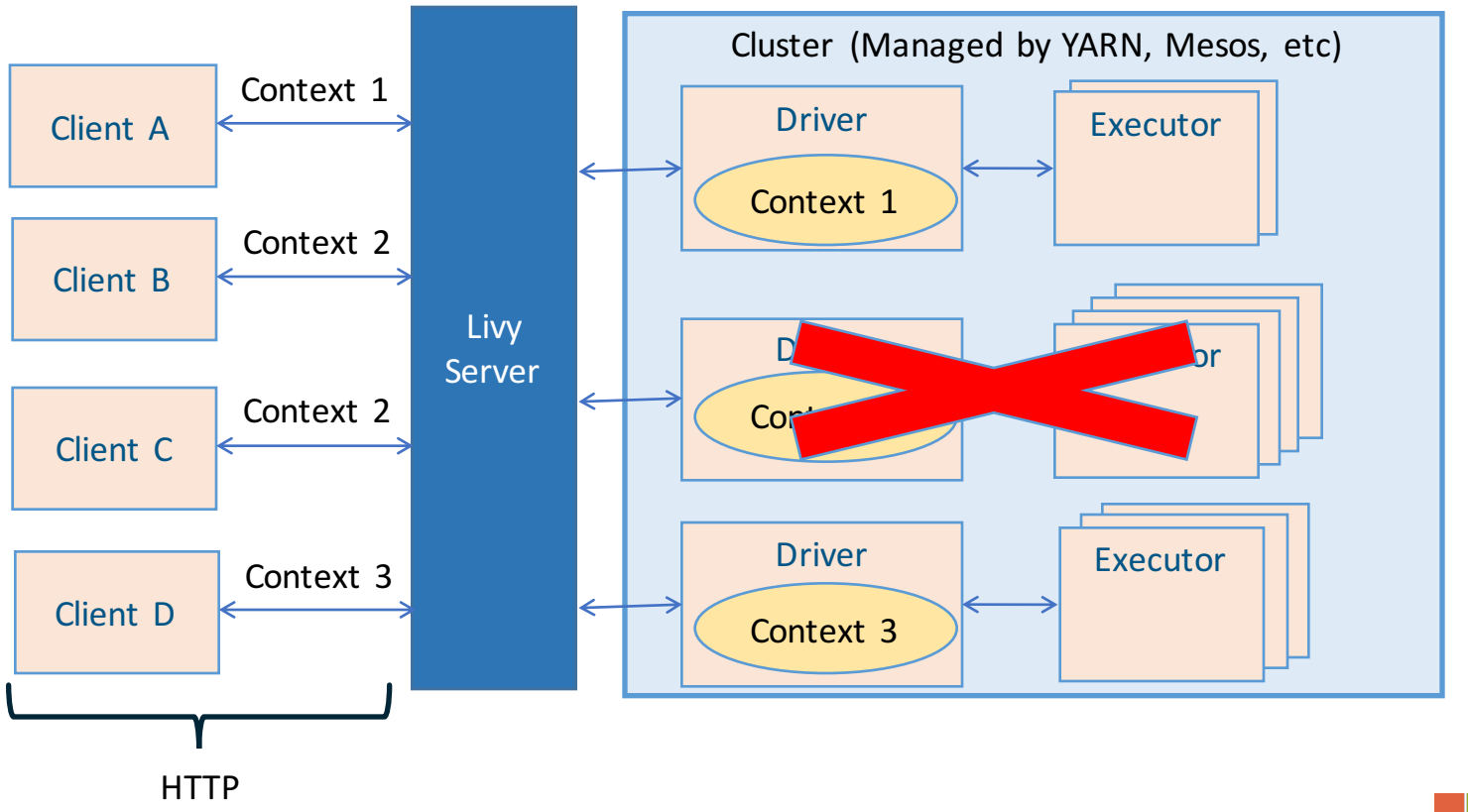
Microsoft

Livy Architecture Highlights

cloudera

Microsoft

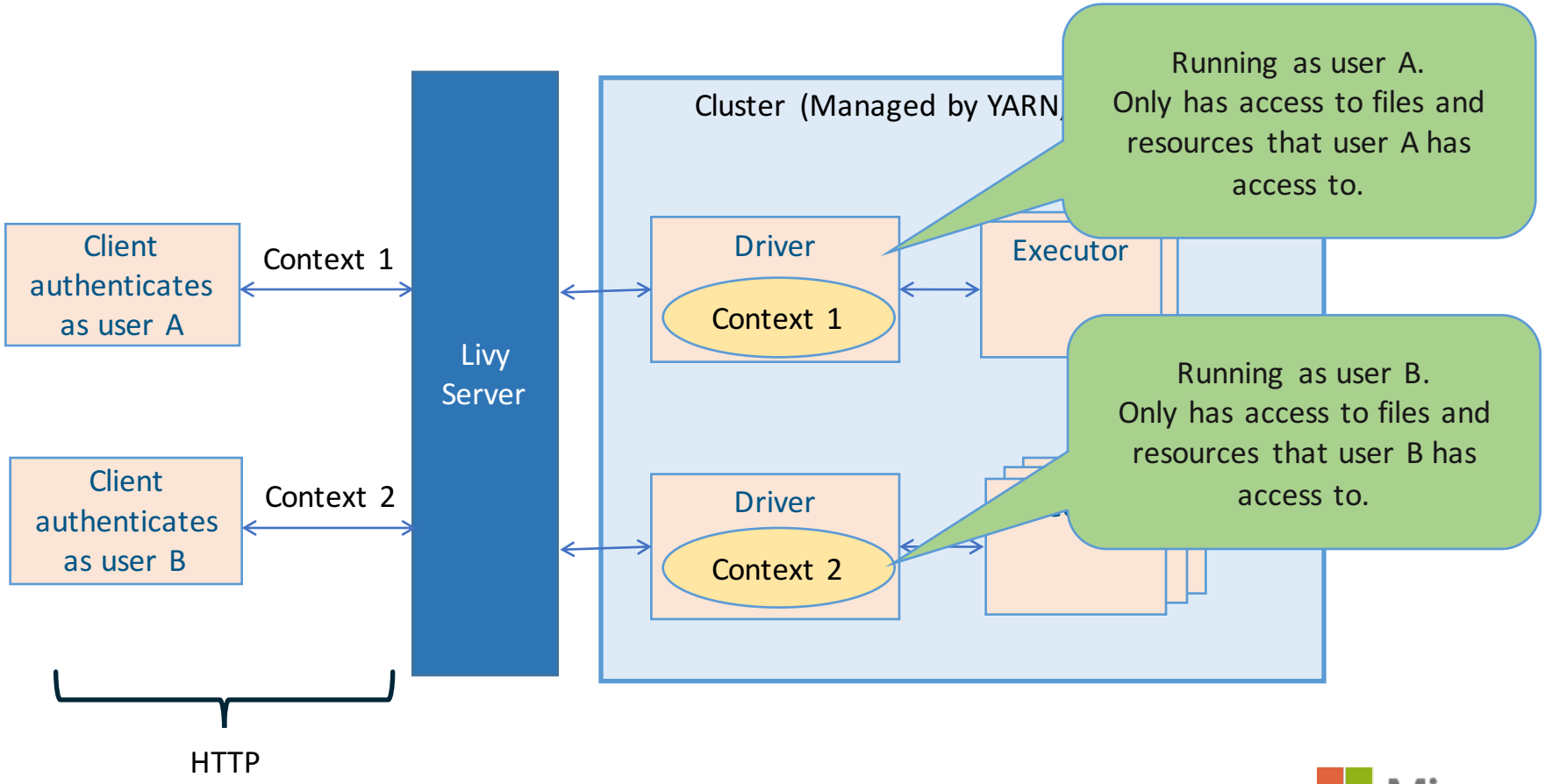
Manage multiple independent Spark Contexts



cloudera

Microsoft

User Impersonation



cloudera

Microsoft

Livy Client API

cloudera

Microsoft

Client API: Job Interface

```
// Job Interface to Implement
public interface Job<T> extends Serializable {
  T call(JobContext jc) throws Exception;
}
```

cloudera

Microsoft

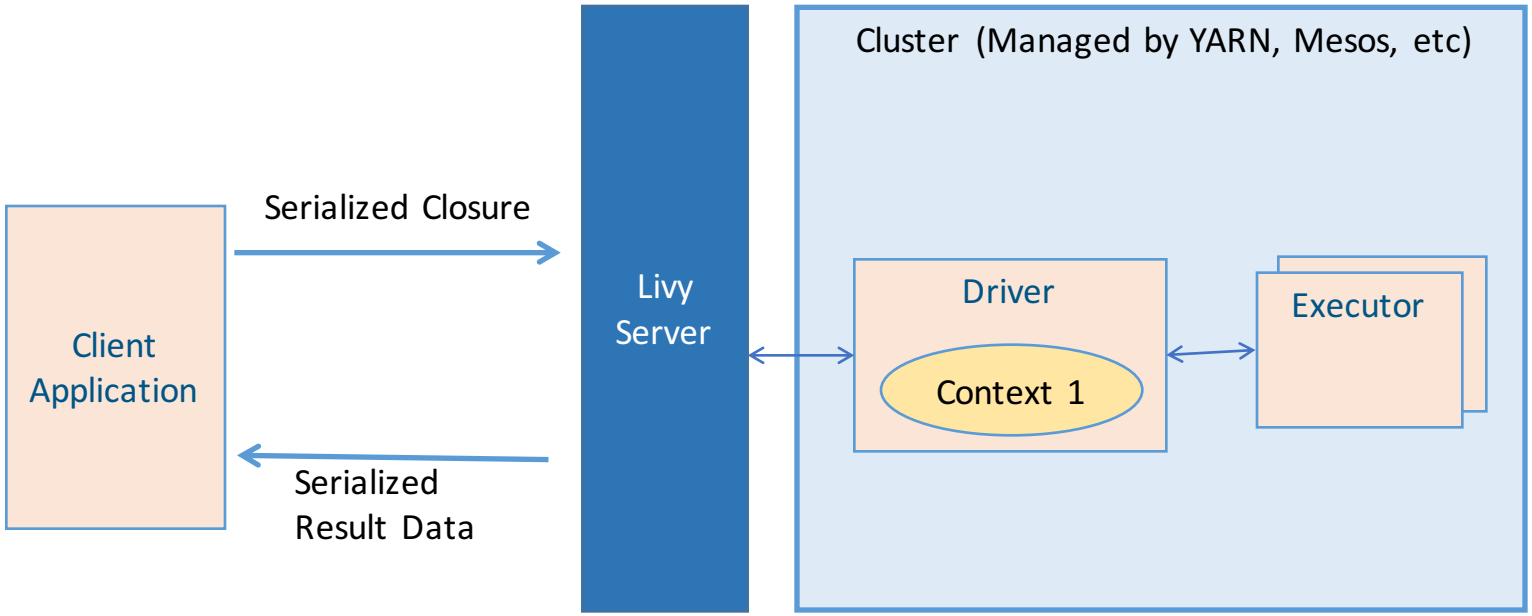
Client API: Submitting a job

```
// Create Livy Client
LivyClient client = new LivyClientBuilder(false)
                          .setURI(new URI("<uri>"))
                          .setAll(<config>
                          .build()

// JobHandle allows monitoring of jobs
JobHandle<Long> handle = client.submit(new YourJob());

// Block until results are returned
Long result = handle.get(TIMEOUT, TimeUnit.SECONDS)

// Close connections
client.stop()
```

cloudera

Microsoft

Client API Architecture



cloudera

Microsoft

Community

http://livy.io

cloudera

Microsoft