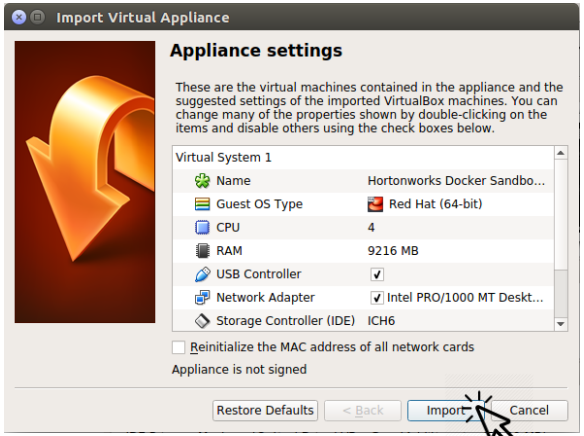
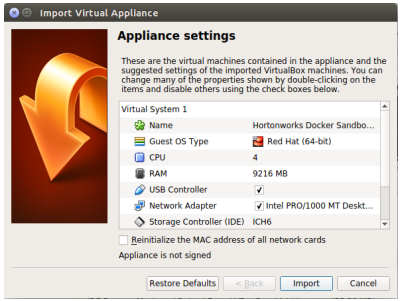

epam
BDCC WORKSHOP
BIG DATA: THE FIRST STEP
September 2018

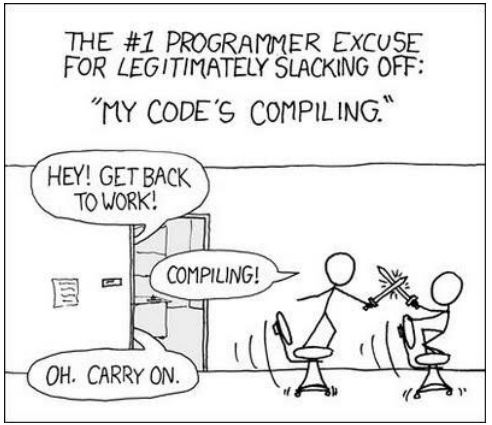
BEFORE START



Open → Import → Start

BEFORE START





Open → Import → Start → ...

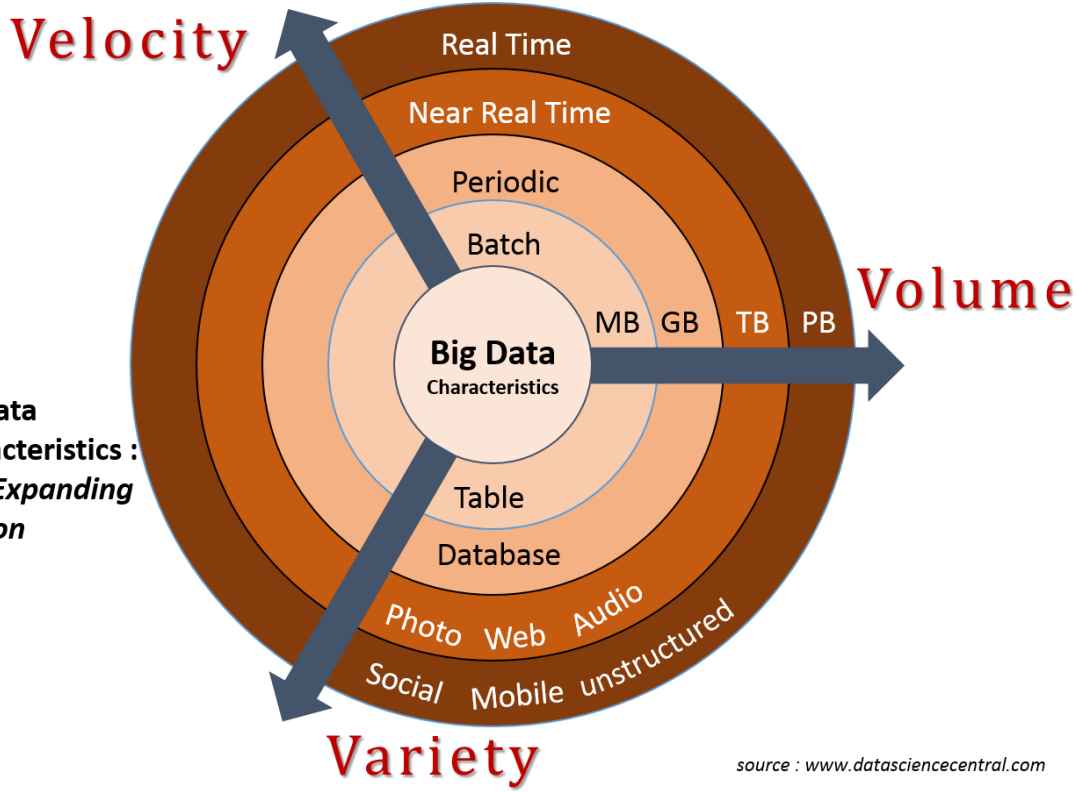

epam
BDCC WORKSHOP
BIG DATA: THE FIRST STEP
September 2018

AGENDA

1. Big Data Overview
2. Case Study
3. Practice

BIG DATA OVERVIEW

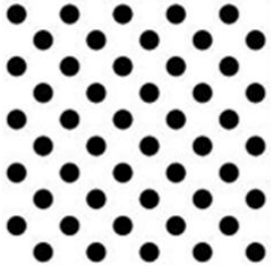
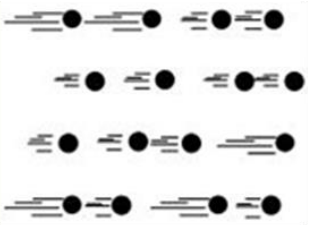
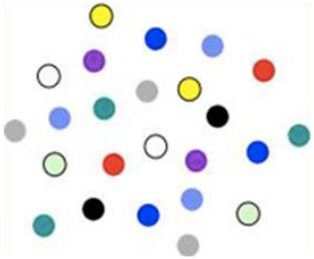
PREREQUISITES #1 (3V)



Big Data Characteristics : ***Ever Expanding horizon***

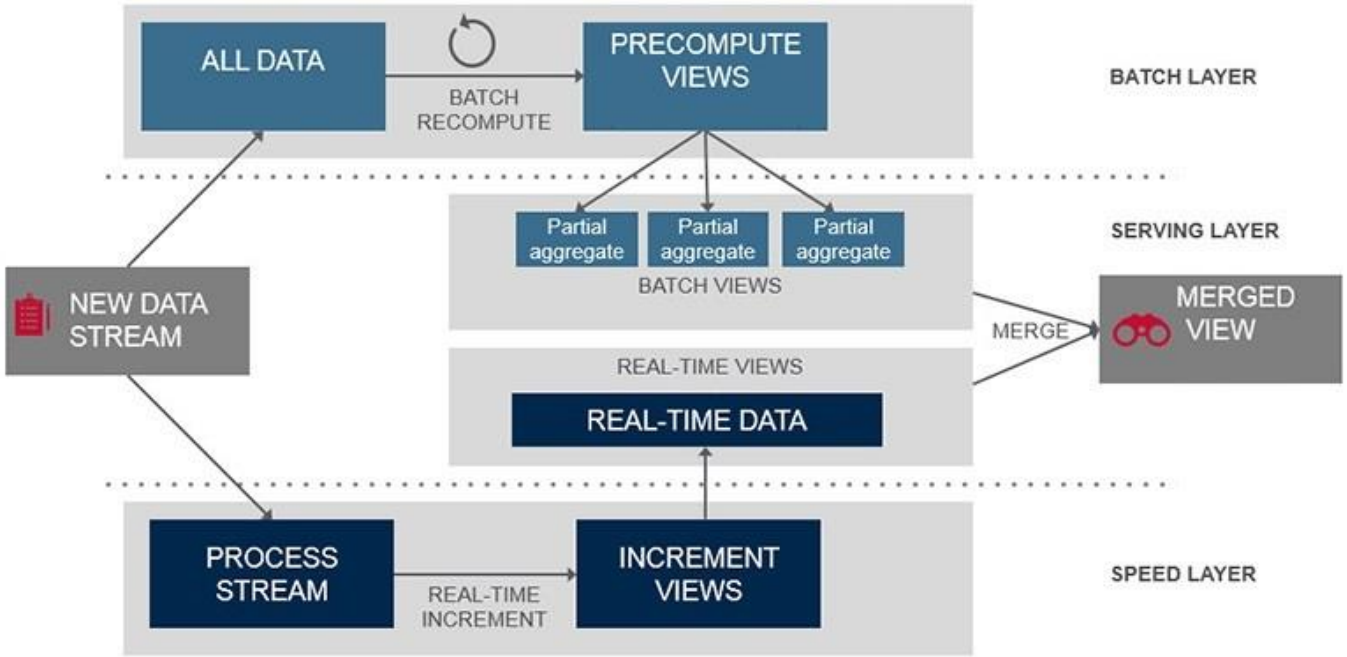
source : www.datasciencecentral.com

PREREQUISITES #2 (5V)

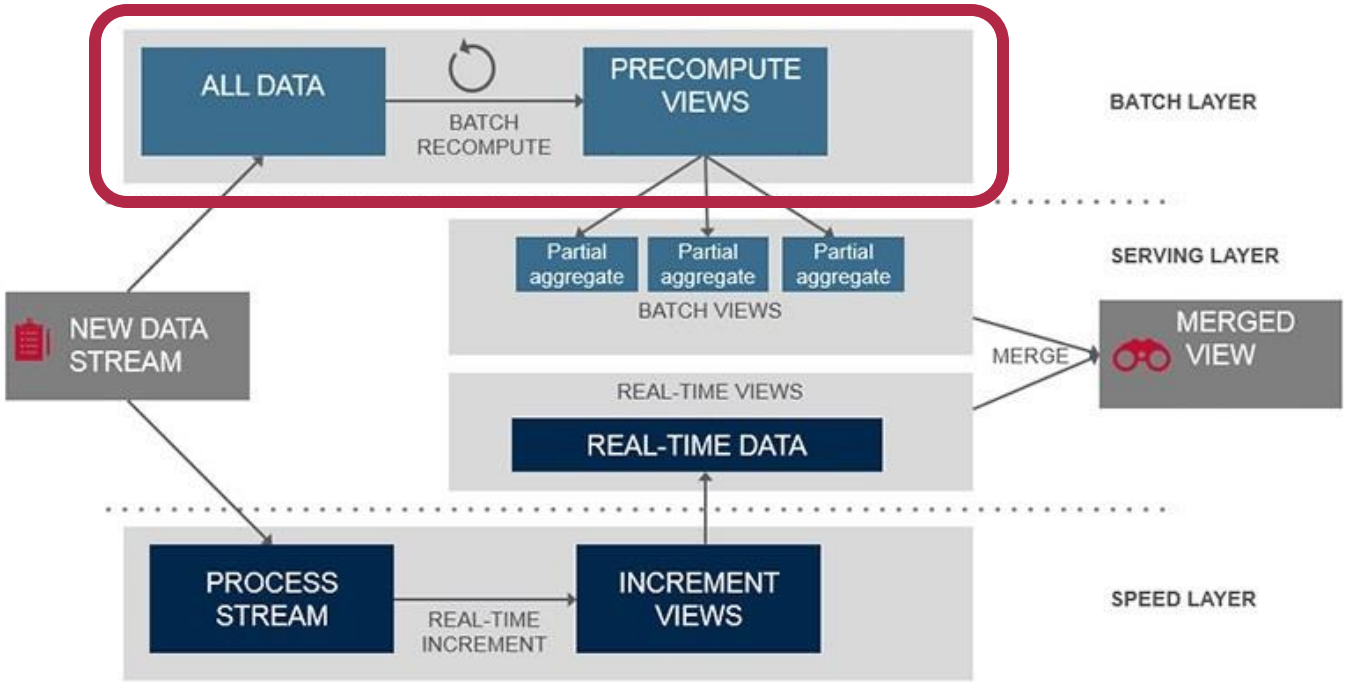
Volume	Velocity	Variety	Veracity	Value
				
Data at Rest	**Data in Motion**	**Data in Many Forms**	**Data in Doubt**	**Data into Money**
Terabytes to Exabytes of existing data to process	Streaming data, requiring milliseconds to seconds to respond	Structured, unstructured, text, multimedia,…	Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations	Business models can be associated to the data

Adapted by a post of Michael Walker on 28 November 2012

BIG DATA APPROACH: LAMBDA ARCHITECTURE



BIG DATA APPROACH: LAMBDA ARCHITECTURE



CASE STUDY

WAREHOUSE MANAGEMENT SERVICE

PROBLEM CONTEXT

BUSINESS ITEMS TO CONSIDER

- Member Application Solution shall support a large number of organizations, each with its own members, business partners, customizations, preferences and application usage specifics.
- All data updates have to pass review/approve lifecycles before being actually applied.
- *The problem of calculating fair commissions to be charged from tenants, based on their shares in the system utilization.*

TECHNICAL ITEMS TO CONSIDER

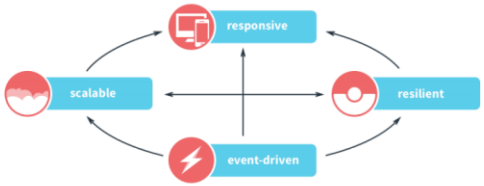
- The solution has to be cloud-ready, cost and resource effective, highly-available (99.9%), fault-tolerant, scalable, performant and ensuring tenant data isolation and protection.
- The solution architecture has to be sustainable, well-integrated with the existing X-Customer's enterprise architecture and comply with X-Customer's policies and standards.
- Not only human-beings, but also automated 3'rd party systems shall be using the platform.

SOLUTION PRESENTATION: OVERVIEW

The solution is inspired by:

- SOA

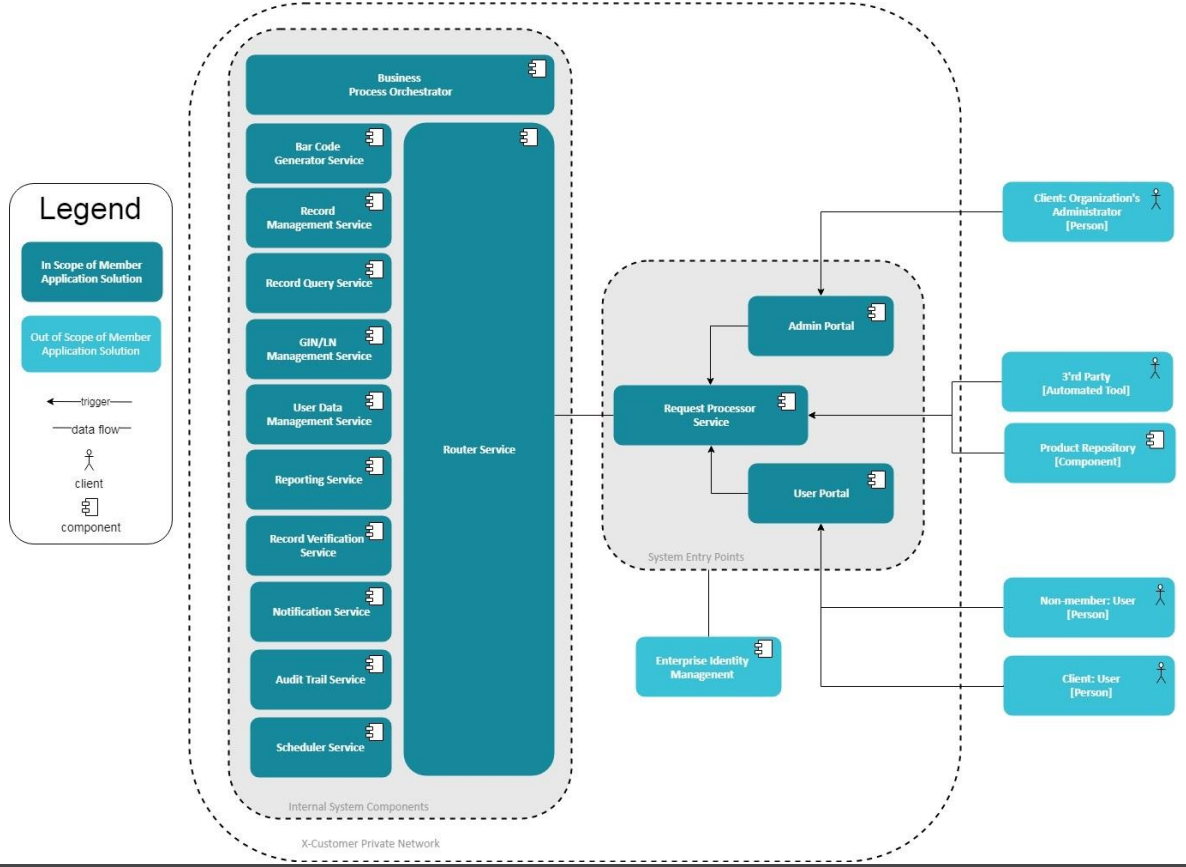
- Reactive Architecture



- Decoupling Reads and Writes

CQRS

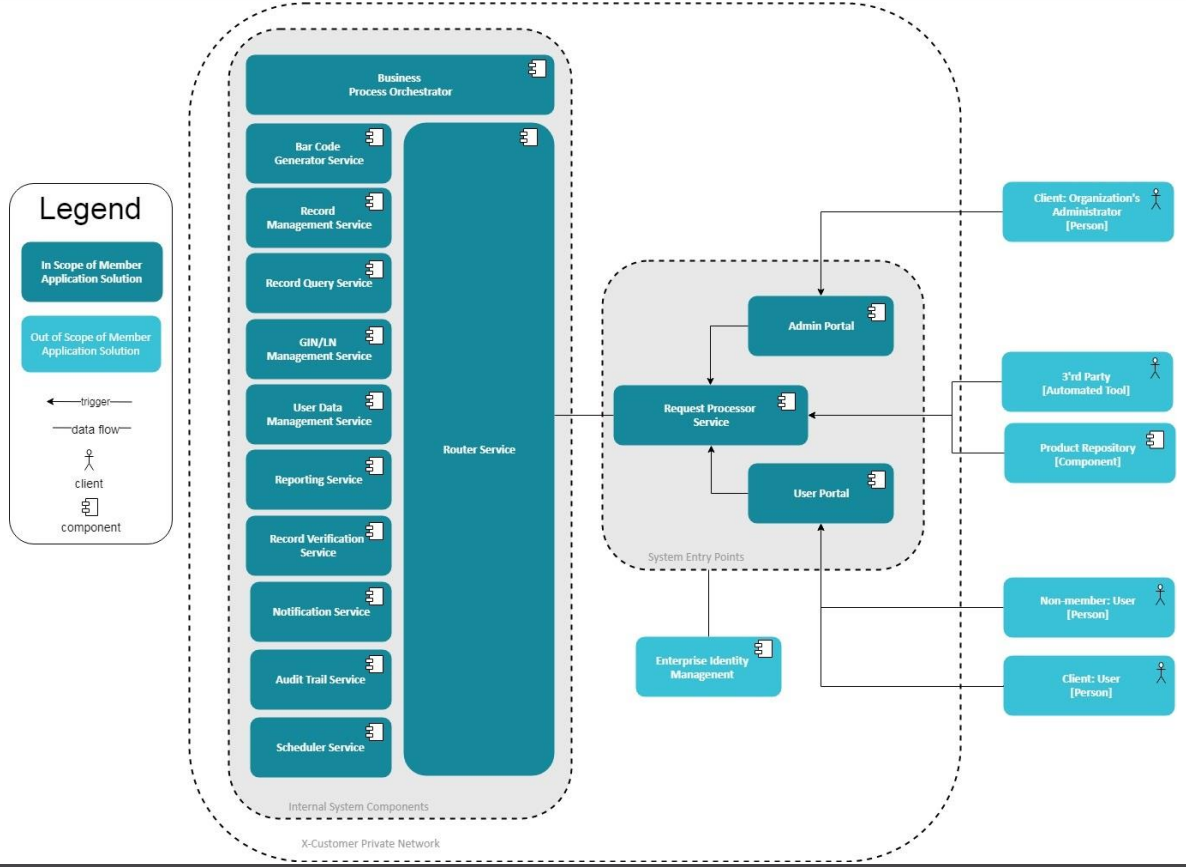
Command Query Responsibility Segregation



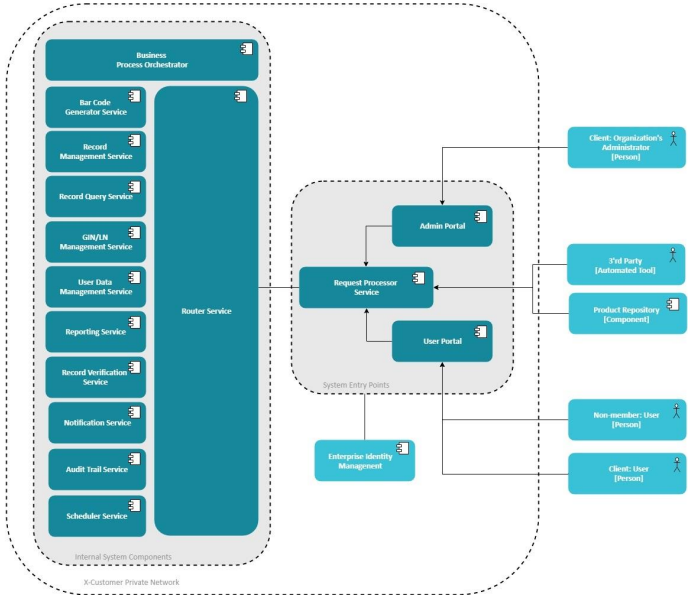
SOLUTION PRESENTATION: RATIONALE

BENEFITS

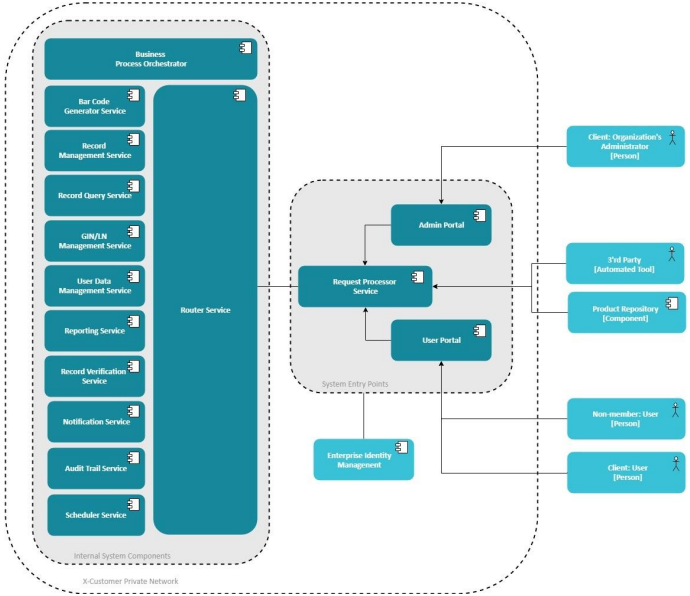
- Scalability
- High-availability
- Security
- Performance
- Fault-tolerance
- Flexibility



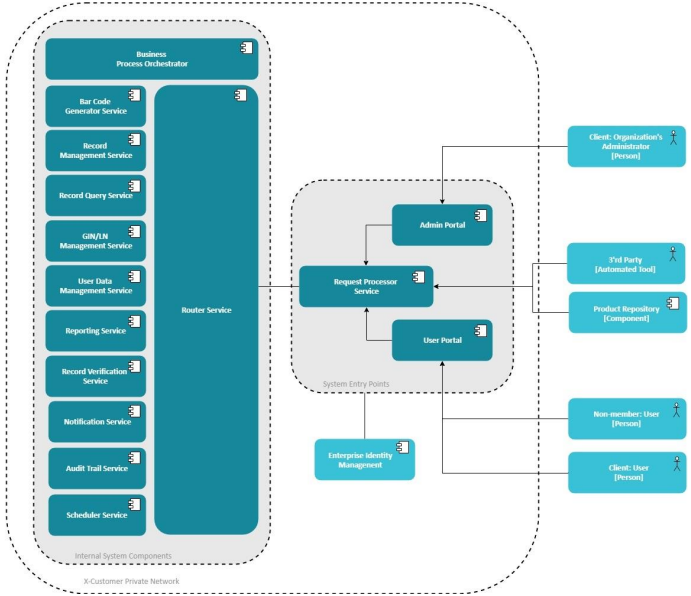
SOLUTION PRESENTATION: FOOTPRINT



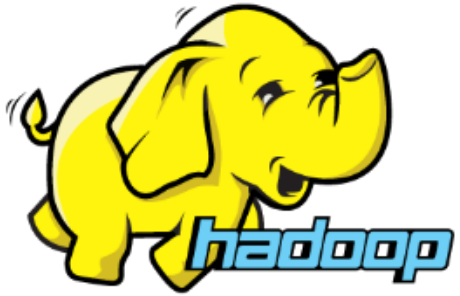
SOLUTION PRESENTATION: FOOTPRINT



SOLUTION PRESENTATION: FOOTPRINT



transactions

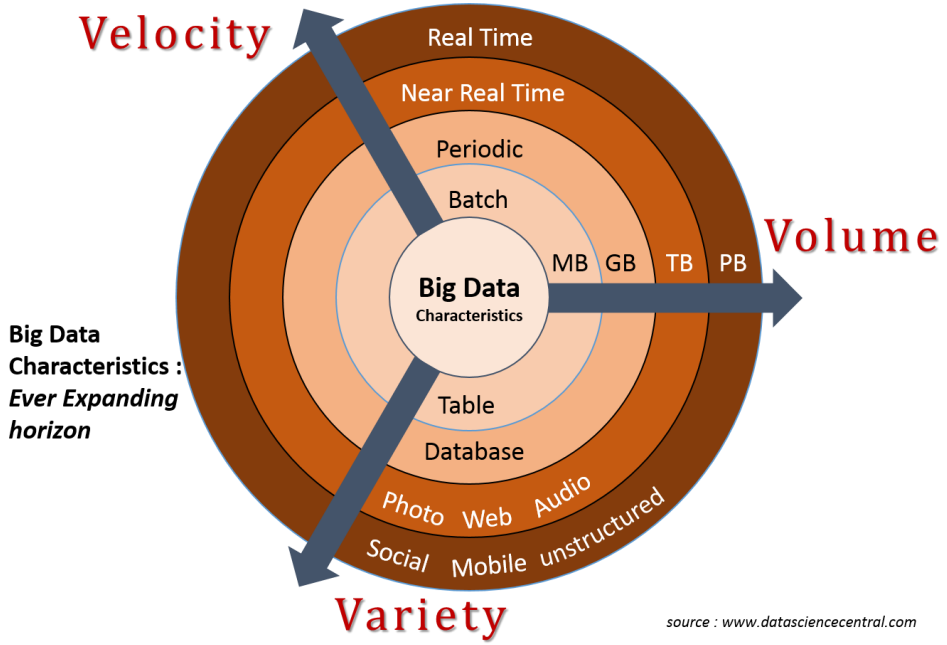


CODE GOES TO DATA

HDFS… AND HERE BIG DATA BEGINS…

SCOPE

- Size of **DATA** >> Size of **Code**



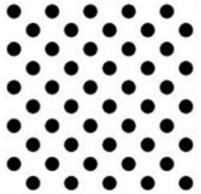
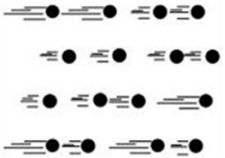
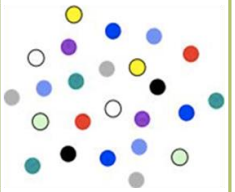
Big Data Characteristics : ***Ever Expanding horizon***

source : www.datasciencecentral.com

HDFS… AND HERE BIG DATA BEGINS…

SCOPE

- Size of DATA >> Size of Code
- $$$ of DATA >> $$$ of HW

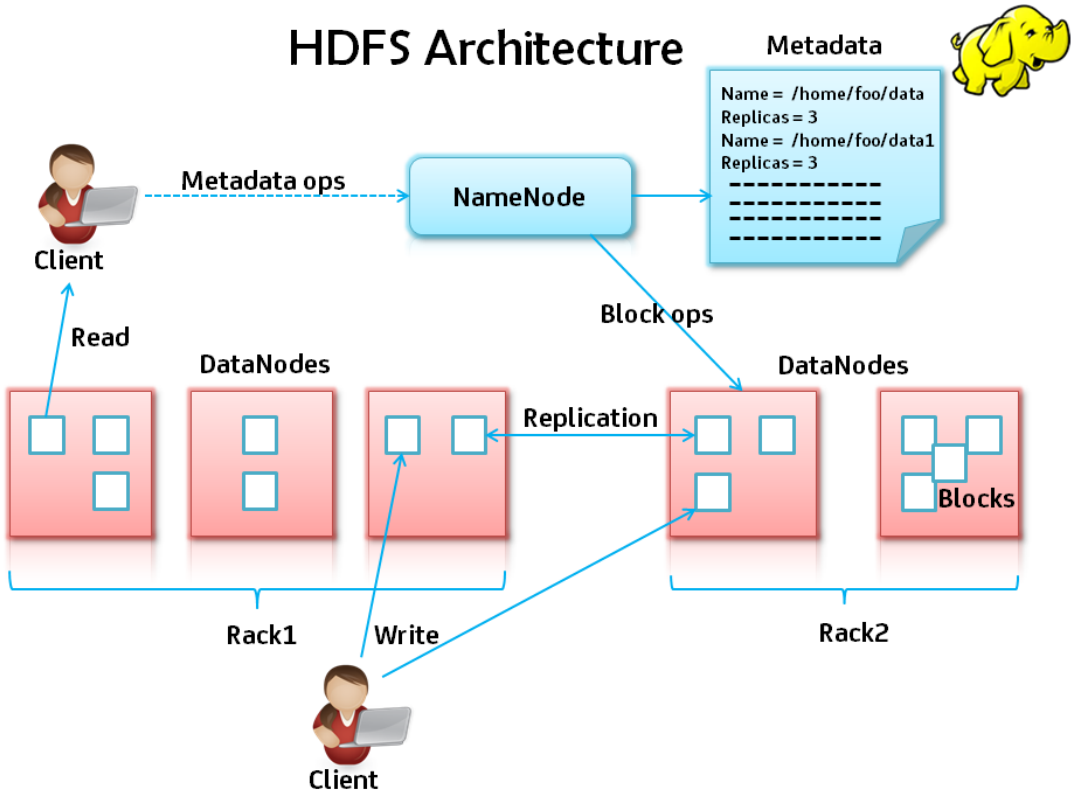
Volume	Velocity	Variety	Veracity	Value
				
Data at Rest	**Data in Motion**	**Data in Many Forms**	**Data in Doubt**	**Data into Money**
Terabytes to Exabytes of existing data to process	Streaming data, requiring milliseconds to seconds to respond	Structured, unstructured, text, multimedia,…	Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations	Business models can be associated to the data

Adapted by a post of Michael Walker on 28 November 2012

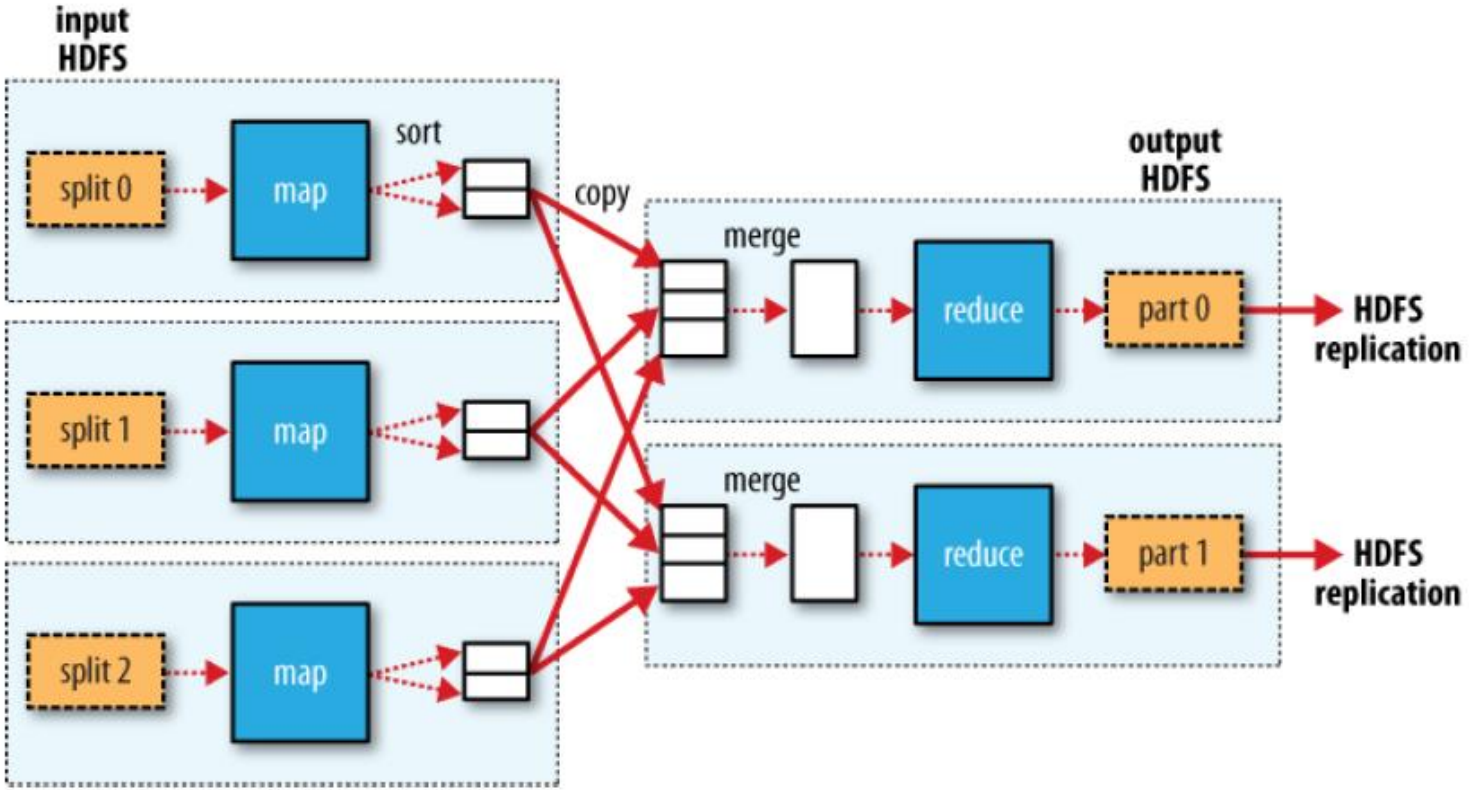
HDFS… AND HERE BIG DATA BEGINS…

PRINCIPLES

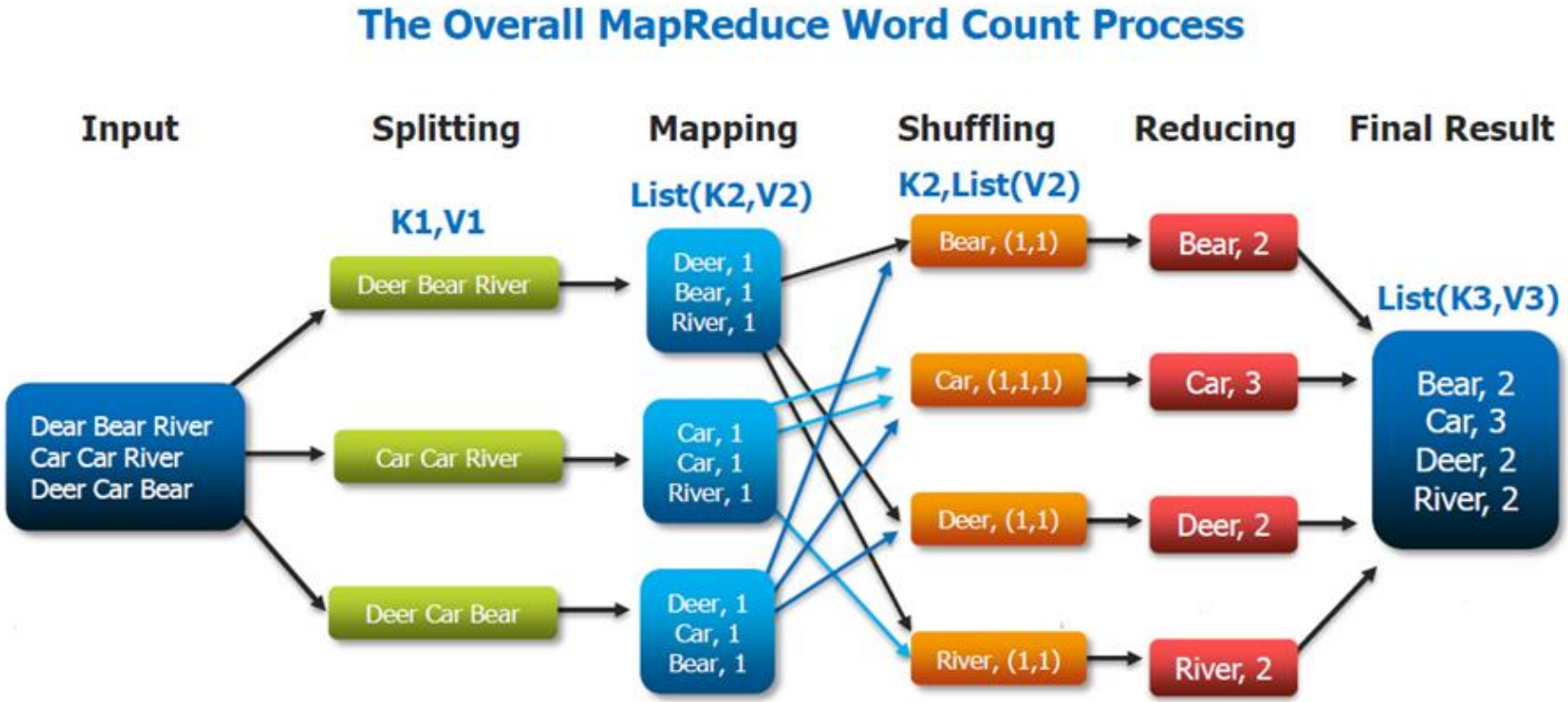
- Spread DATA across HWs
- Copy code across HWs
- Let code work with local data (**only**)



MAP-REDUCE


input
HDFS
split 0
map
sort
split 1
map
split 2
map
copy
merge
reduce
part 0
output
HDFS
HDFS
replication
merge
reduce
part 1
HDFS
replication

MAP-REDUCE


The Overall MapReduce Word Count Process
Input
Splitting
Mapping
Shuffling
Reducing
Final Result
K1,V1
List(K2,V2)
K2,List(V2)
List(K3,V3)
Dear Bear River
Car Car River
Deer Car Bear
Deer Bear River
Car Car River
Deer Car Bear
Deer, 1
Bear, 1
River, 1
Car, 1
Car, 1
River, 1
Deer, 1
Car, 1
Bear, 1
Bear, (1,1)
Car, (1,1,1)
Deer, (1,1)
River, (1,1)
Bear, 2
Car, 3
Deer, 2
River, 2
Bear, 2
Car, 3
Deer, 2
River, 2

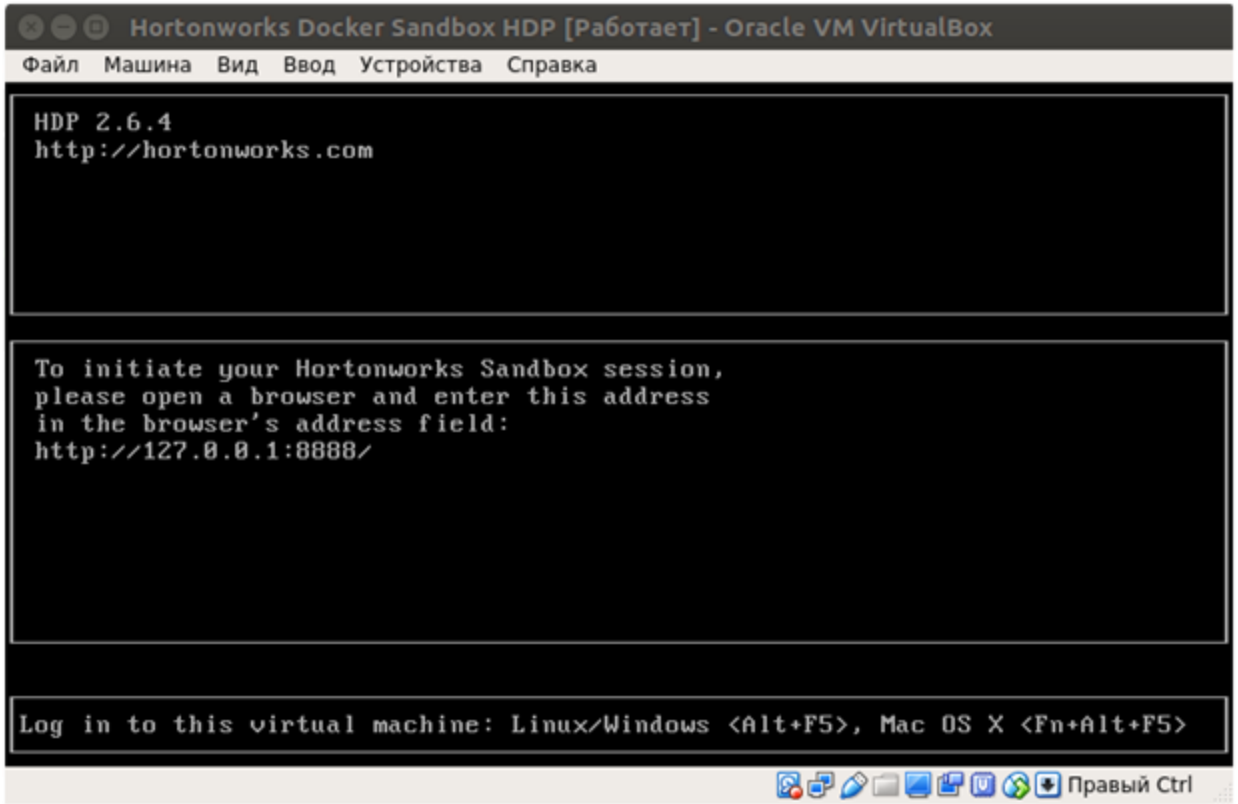
BIG DATA LANDSCAPE

BIG DATA LANDSCAPE

BIG DATA LANDSCAPE

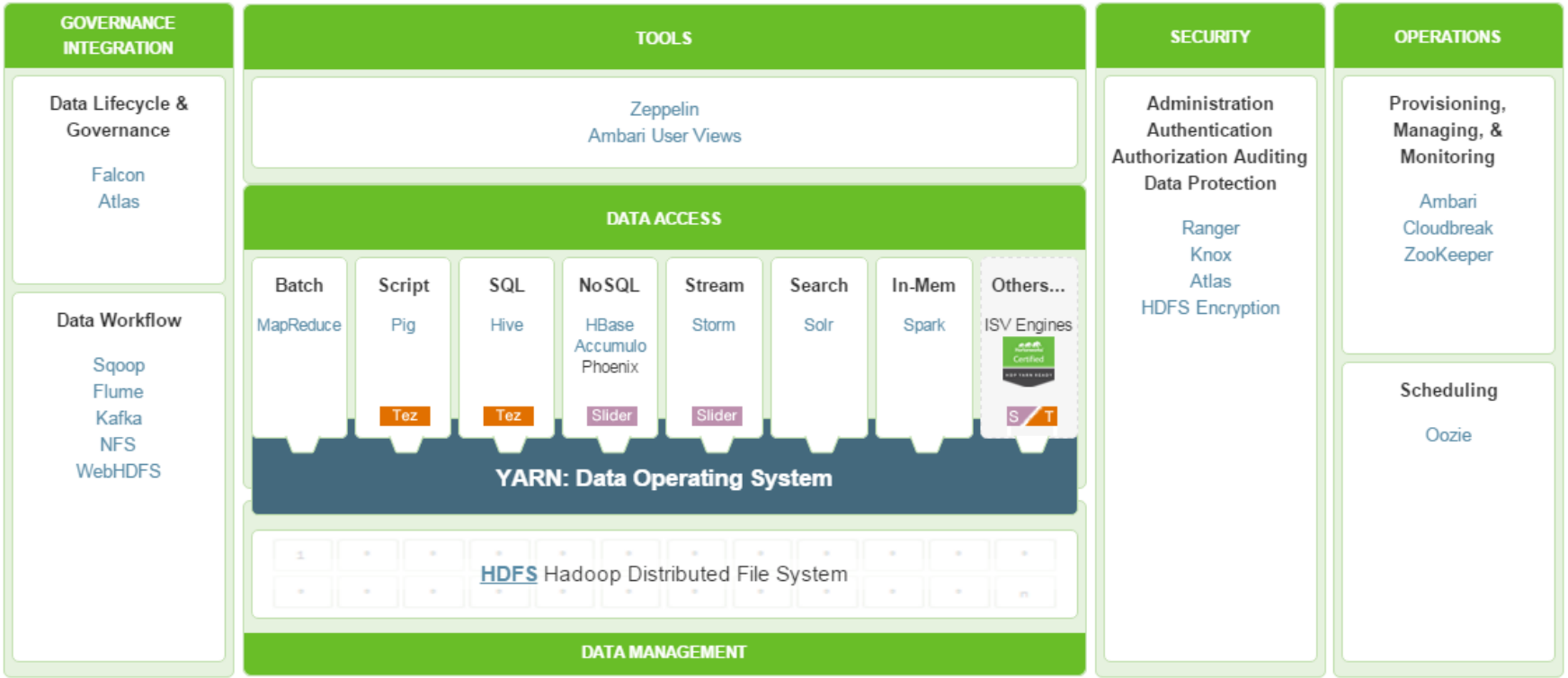

Coffee
Time!

CHECK YOURSELF

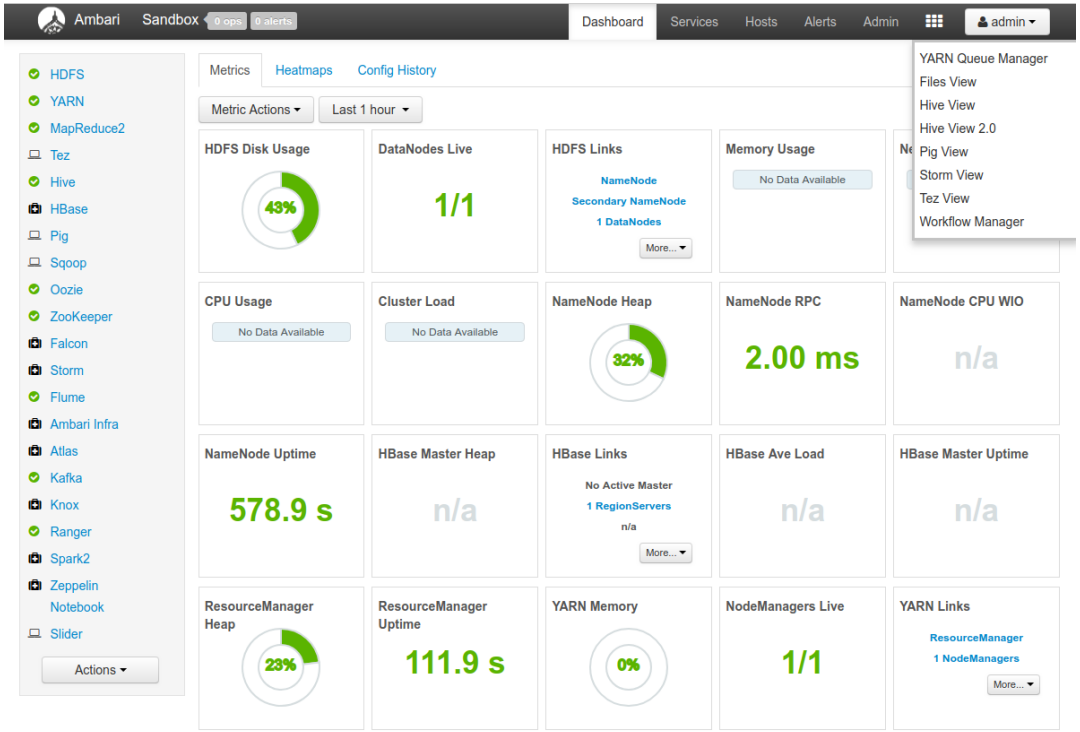


HORTONWORKS

HORTONWORKS



LET'S PLAY



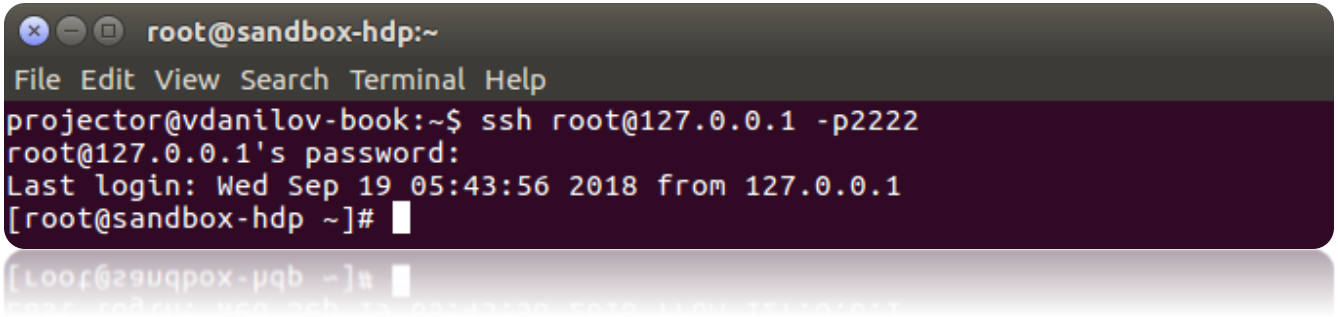
- Login

  address: http://127.0.0.1:8080/

  creds: **admin / workshop**

http://127.0.0.1:8080

LET'S PLAY



- SSH

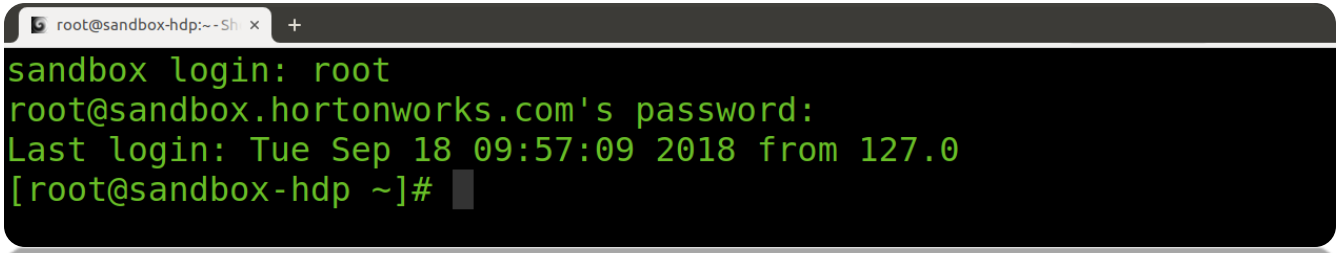
  ```
  # ssh root@127.0.0.1 -p2222
  ```

  creds: **root / workshop**

- Run

  ```
  # cd workshop
  # ls -la
  ```

LET'S PLAY



- Login

  address: http://127.0.0.1:4200/ , creds: **root / workshop**

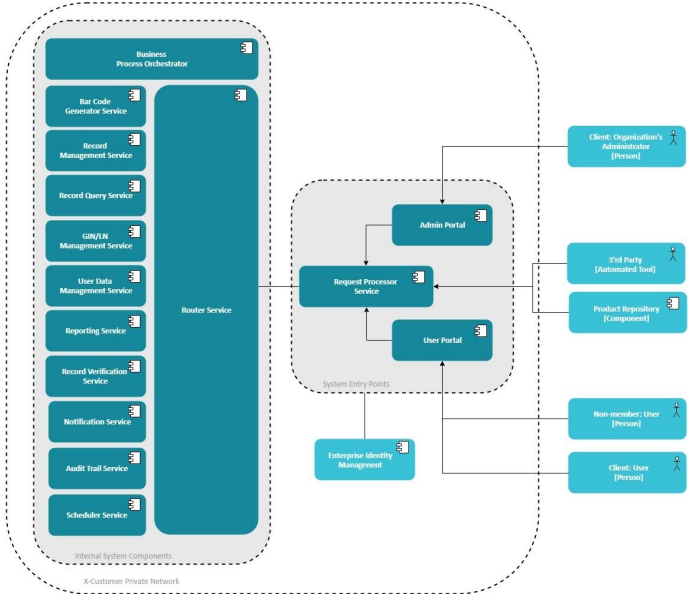
- Run

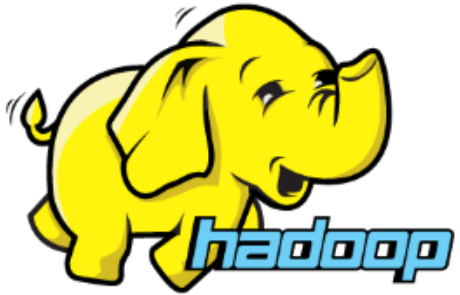
```
# cd workshop
# ls -la
```

THE APPLICATION

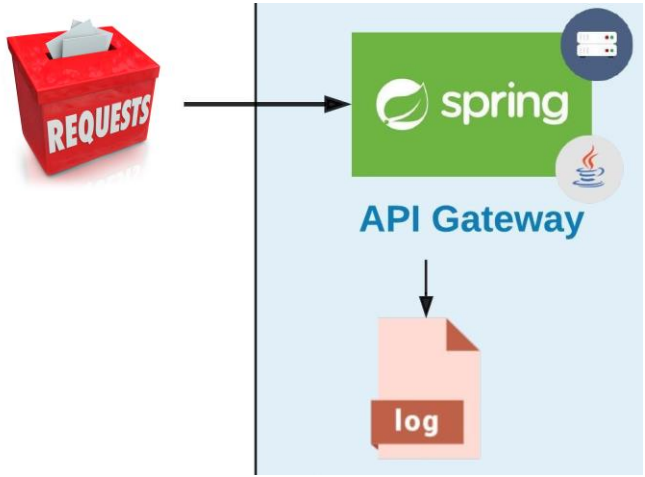
HTTPS://GITHUB.COM/EPAM-ZHUJ/BIGDATA-WORKSHOP

SOLUTION PRESENTATION: FOOTPRINT

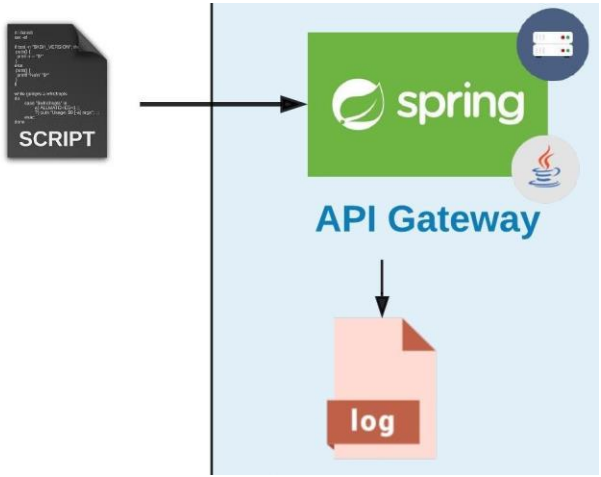




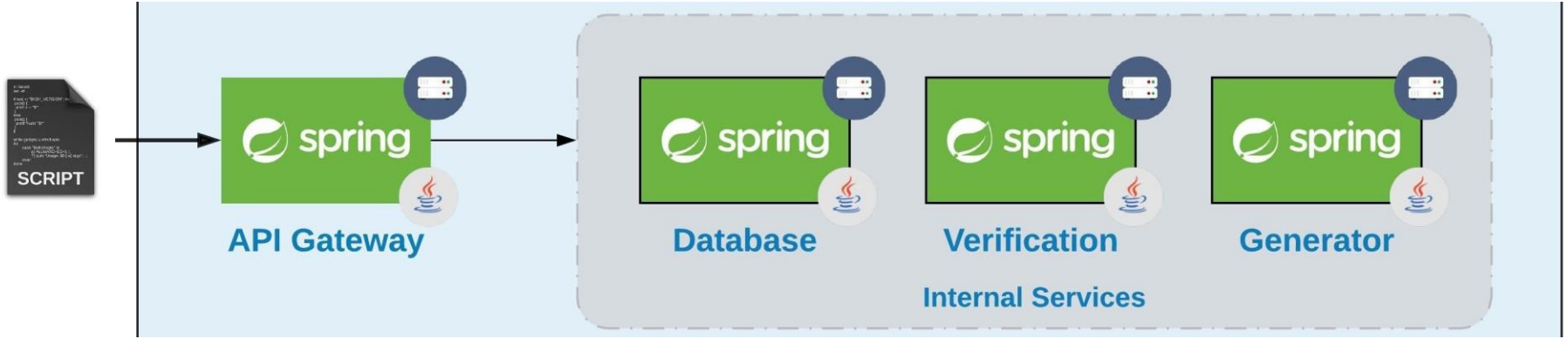
PUBLIC ENDPOINT



PUBLIC ENDPOINT



LET’S INTRODUCE A COUPLE OF SERVICES

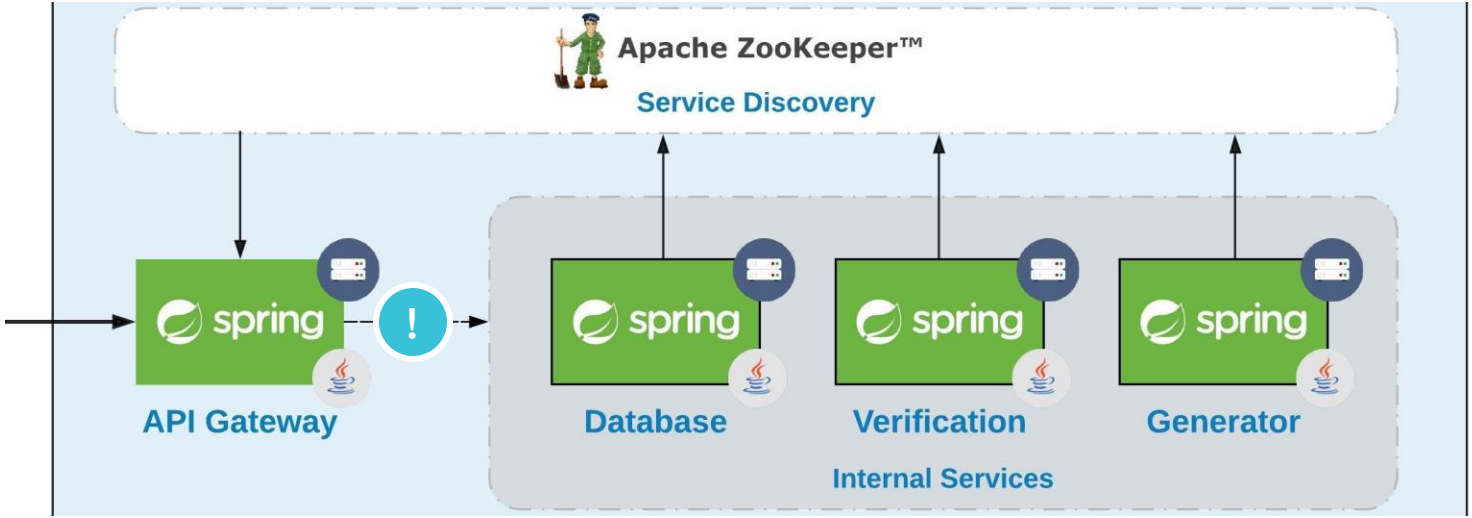


PUBLIC ENDPOINT

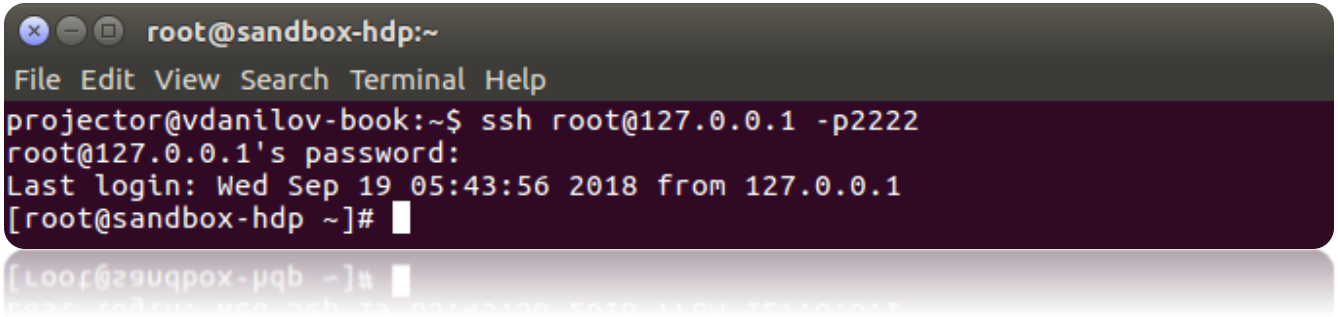

```
{
  "userIds": ["Dmitrii", "Jennifer", "Daniel", "Olga", "Kathryn", "Sunil", "Maria", "Alexander"],
  "numberOfRequests": 1000,
  "workflows": [
    {
      "steps": [
        {"serviceName": "database", "invocationDelay": 0},
        {"serviceName": "verification", "invocationDelay": 0},
        {"serviceName": "generator", "invocationDelay": 0}
      ]
    },
    {
      "steps": [
        {"serviceName": "database", "invocationDelay": 1000},
        {"serviceName": "verification", "invocationDelay": 1000},
        {"serviceName": "generator", "invocationDelay": 1000}
      ]
    }
  ]
}
```

solution-services/api-gateway/src/main/resources/request_template.json

SERVICE DISCOVERY TO THE RESCUE



LET'S PLAY



- Login

  ```
  # ssh root@127.0.0.1 -p2222
  ```

  creds: **root / workshop**

- Run

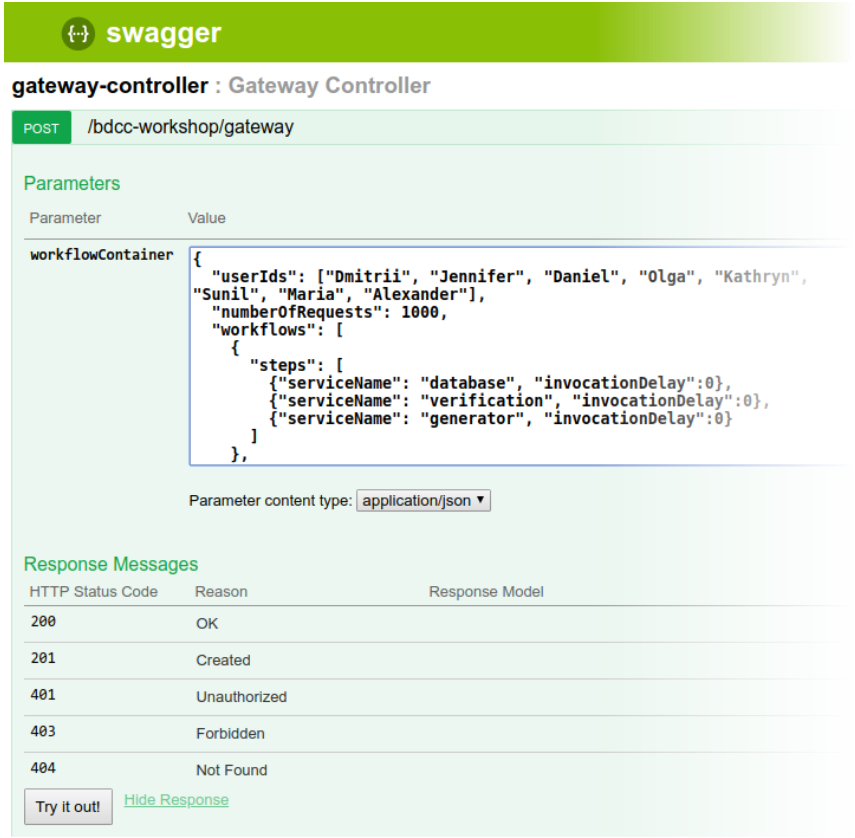
  ```
  # cd workshop/api-gateway
  # java -jar api-gateway-1.0-SNAPSHOT.jar
  ```

- Run other services

http://127.0.0.1:4200/

```
ssh root@127.0.0.1 -p2222
```

LET'S PLAY

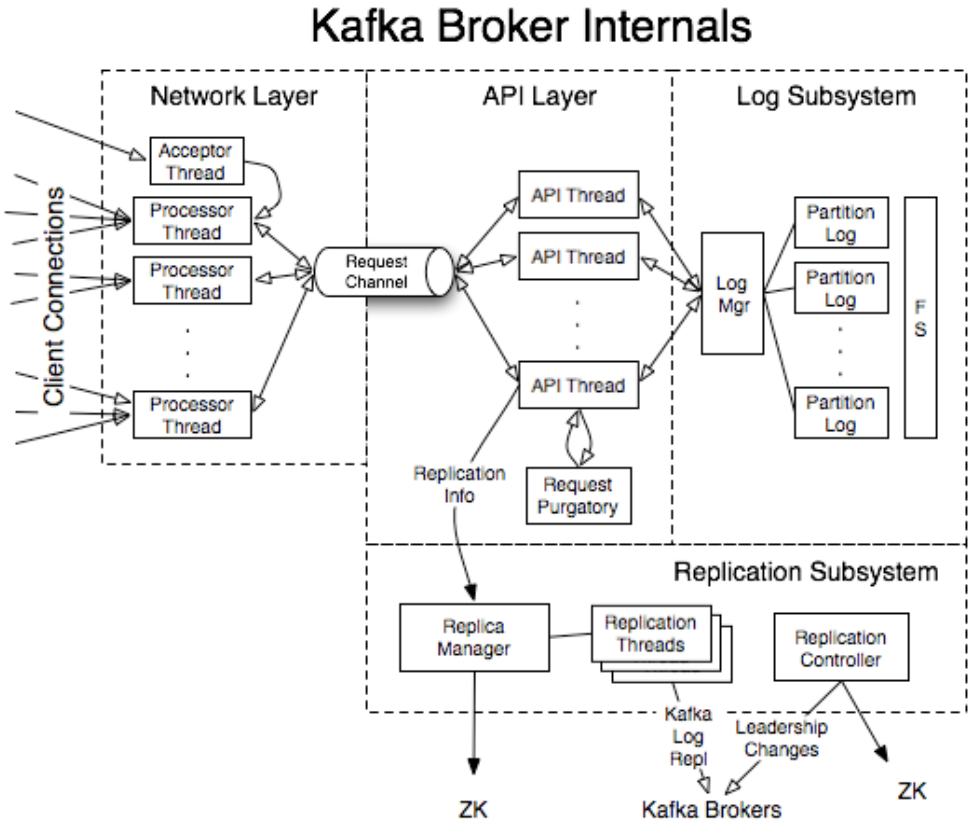


- Open http://127.0.0.1:15505/swagger-ui.html
- Find endpoint / method
  endpoint = gateway-controller / POST
- Set script text
  solution-services/.../request_template.json
- Try it out!

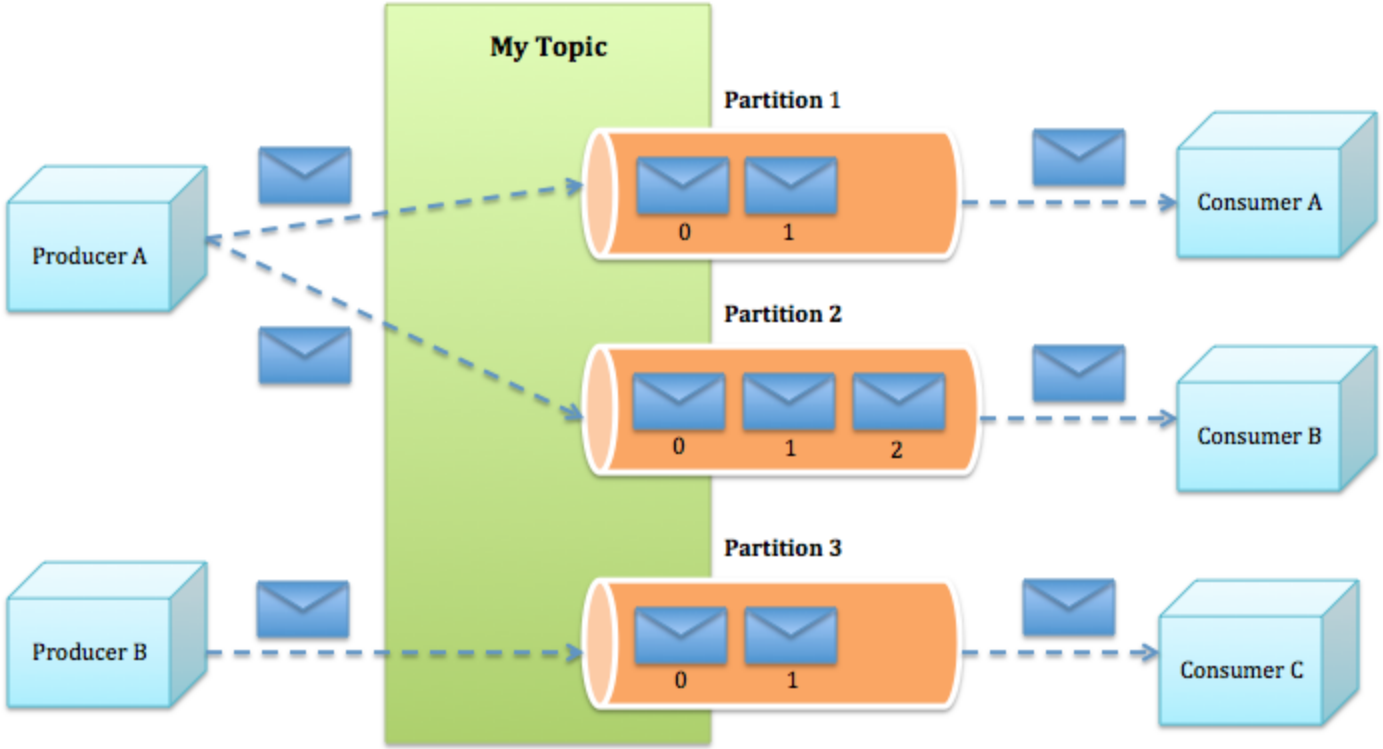
http://127.0.0.1:15505/swagger-ui.html

APACHE KAFKA

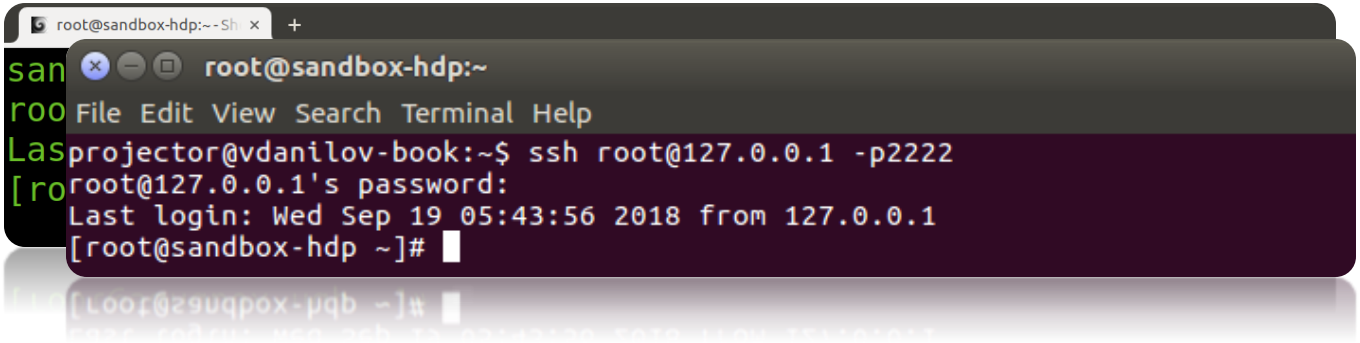
APACHE KAFKA



APACHE KAFKA



LET'S PLAY



- Login twice

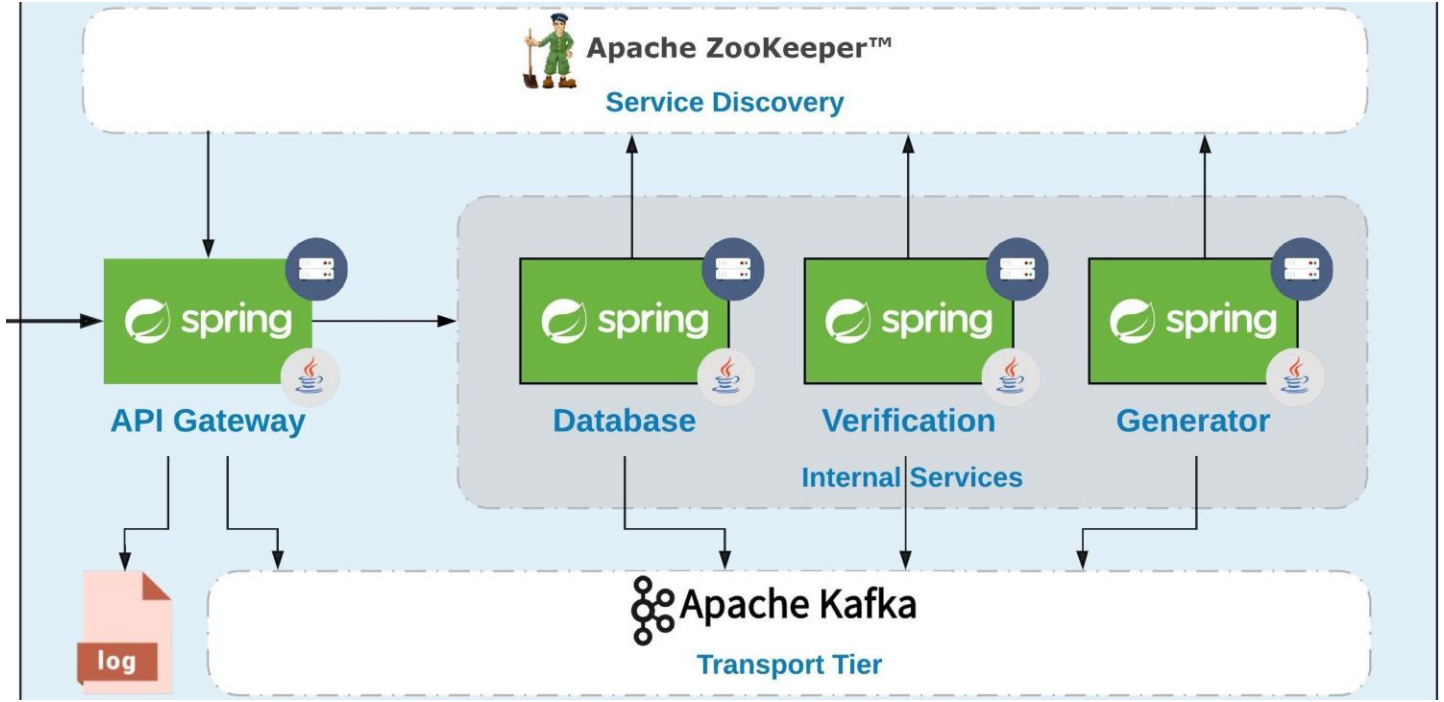
  address: http://127.0.0.1:4200/ , creds: **root / workshop**

- Run

```
(*)# cd cd workshop/kafka
(1)# ./kafka-console-producer.sh --topic test --broker-list 0.0.0.0:6667
(2)# ./kafka-console-consumer.sh --topic test --bootstrap-server 0.0.0.0:6667
```

- Have fun

APPLICATION



MESSAGES

api-gateway

```
"[{userId}] [{workflowId}] [{step.serviceName}]" -> log
"{userId}" -> gateway.user.activity.topic
```

database-service

```
"[database] {ts} | {userId} | {workflowId} | putSize: {szInKb}; returnSize: {szIbKb}"
  -> resource.utilization.topic
```

verification-service

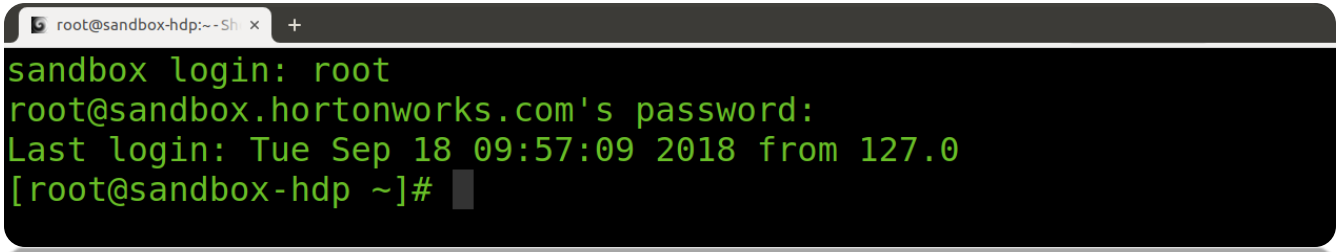
```
"[verification] {ts} | {userId} | {workflowId} | records verified: {number}"
  -> resource.utilization.topic
```

generator-service

```
"[generator] ({ts1}, {ts2}) | {userId} | {workflowId} | avg cpu time: {avgCpuTime}"
  -> resource.utilization.topic
```

workflowId = UUID.*randomUUID*(), ts = yyyy-MM-dd HH:mm:ss.SSS

LET'S PLAY



- Login

  address: http://127.0.0.1:4200/ , creds: **root / workshop**

- Run

```
# cd workshop/kafka
# ./kafka-console-consumer.sh --topic resource.utilization.topic
                              --bootstrap-server 0.0.0.0:6667
```

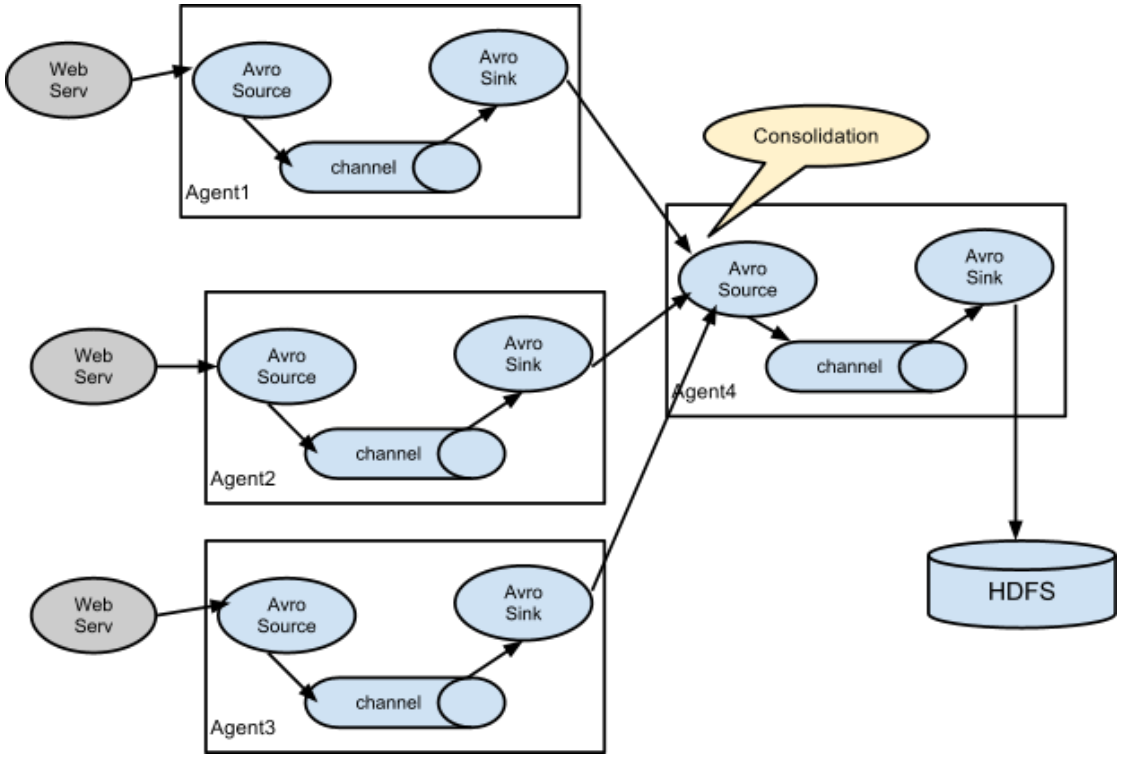
- Have fun (with the App)

LET'S PLAY

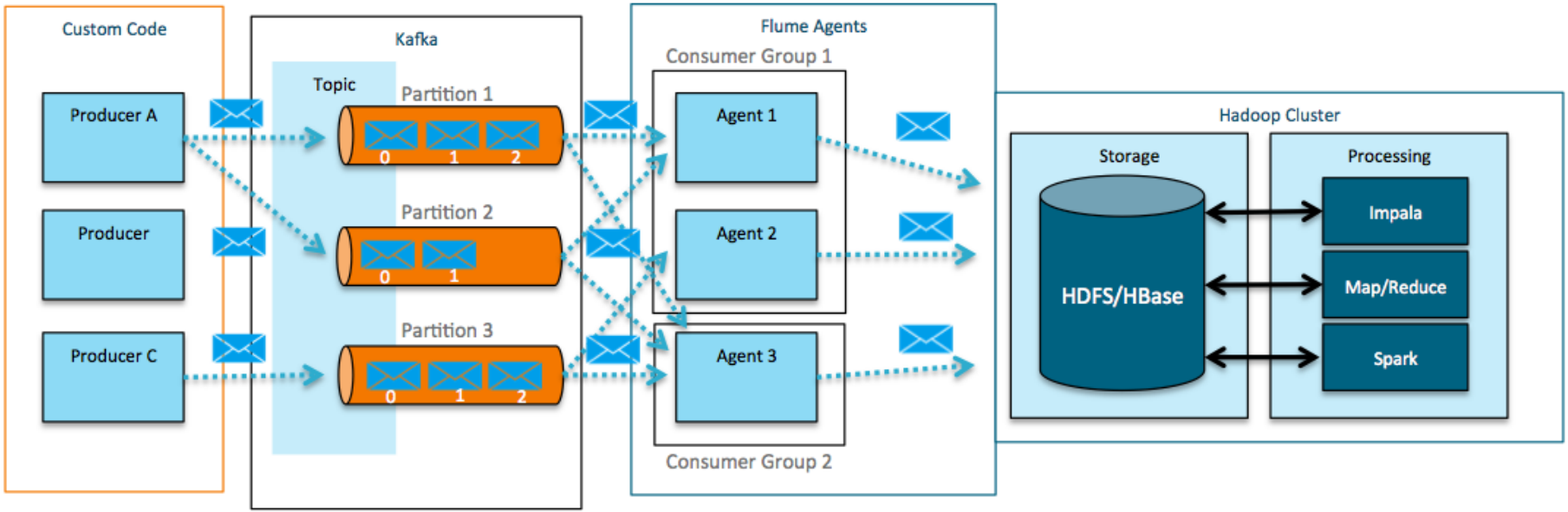

generator] (2018-09-18 18:53:08.569, 2018-09-18 18:53:11.403) | Alexander | aa9d362e-92e5-4034-b05a-ca8991932eba | avg cpu time: 0.6476321919684
generator] (2018-09-18 18:53:08.570, 2018-09-18 18:53:12.283) | Olga | 4c43b98a-4c0d-4558-9516-610994e96067 | avg cpu time: 0.14325644632534296
generator] (2018-09-18 18:53:08.575, 2018-09-18 18:53:10.114) | Olga | be9a656d-8921-4fbb-95ec-866041a9715b | avg cpu time: 0.5525513526276564
database] 2018-09-18 18:53:08.591 | Jennifer | 81352f88-3064-4945-becb-769328af4876 | putSize: 298; returnSize: 1617
database] 2018-09-18 18:53:08.606 | Daniel | 10a85783-0c8b-48f9-b4ca-2ff4db7109ac | putSize: 877; returnSize: 1153
database] 2018-09-18 18:53:08.627 | Daniel | 5e40d661-efca-4938-825f-70b2131518ce | putSize: 113; returnSize: 1793
database] 2018-09-18 18:53:08.628 | Sunil | 00605498-e10c-44f3-a935-47ed1824fe4b | putSize: 529; returnSize: 2518
verification] 2018-09-18 18:53:08.653 | Jennifer | 81352f88-3064-4945-becb-769328af4876 | records verified: 26233
verification] 2018-09-18 18:53:08.668 | Daniel | 10a85783-0c8b-48f9-b4ca-2ff4db7109ac | records verified: 27517
verification] 2018-09-18 18:53:08.674 | Sunil | 00605498-e10c-44f3-a935-47ed1824fe4b | records verified: 4514
verification] 2018-09-18 18:53:08.679 | Daniel | 5e40d661-efca-4938-825f-70b2131518ce | records verified: 43605
generator] (2018-09-18 18:53:08.692, 2018-09-18 18:53:18.125) | Jennifer | 81352f88-3064-4945-becb-769328af4876 | avg cpu time: 0.1442643634575
generator] (2018-09-18 18:53:08.700, 2018-09-18 18:53:15.273) | Daniel | 10a85783-0c8b-48f9-b4ca-2ff4db7109ac | avg cpu time: 0.7417449185107282
generator] (2018-09-18 18:53:08.716, 2018-09-18 18:53:17.104) | Sunil | 00605498-e10c-44f3-a935-47ed1824fe4b | avg cpu time: 0.5113510081448867
database] 2018-09-18 18:53:08.725 | Dmitrii | 55dbd479-17da-4942-ad35-b0a9fb9baa67 | putSize: 695; returnSize: 795
generator] (2018-09-18 18:53:08.734, 2018-09-18 18:53:12.083) | Daniel | 5e40d661-efca-4938-825f-70b2131518ce | avg cpu time: 0.032523965044705
verification] 2018-09-18 18:53:08.759 | Dmitrii | 55dbd479-17da-4942-ad35-b0a9fb9baa67 | records verified: 35498
generator] (2018-09-18 18:53:08.780, 2018-09-18 18:53:16.664) | Dmitrii | 55dbd479-17da-4942-ad35-b0a9fb9baa67 | avg cpu time: 0.421702843676892
database] 2018-09-18 18:53:09.738 | Jennifer | c7292255-b407-4e3d-845d-e5c9609fe59c | putSize: 663; returnSize: 684
database] 2018-09-18 18:53:09.749 | Daniel | c72f3629-4e9b-453f-b06d-bb95547fcdf7 | putSize: 625; returnSize: 1666
database] 2018-09-18 18:53:09.768 | Olga | 24cefdfe-cb1e-4b75-9d6d-5f3f8281d007 | putSize: 933; returnSize: 1585
database] 2018-09-18 18:53:09.824 | Daniel | 001c59a2-d38c-4f28-90ae-8e440bc64101 | putSize: 430; returnSize: 2774
verification] 2018-09-18 18:53:10.767 | Jennifer | c7292255-b407-4e3d-845d-e5c9609fe59c | records verified: 30644
verification] 2018-09-18 18:53:10.782 | Daniel | c72f3629-4e9b-453f-b06d-bb95547fcdf7 | records verified: 17500
verification] 2018-09-18 18:53:10.803 | Olga | 24cefdfe-cb1e-4b75-9d6d-5f3f8281d007 | records verified: 38760
verification] 2018-09-18 18:53:10.840 | Daniel | 001c59a2-d38c-4f28-90ae-8e440bc64101 | records verified: 32422
generator] (2018-09-18 18:53:11.796, 2018-09-18 18:53:17.216) | Jennifer | c7292255-b407-4e3d-845d-e5c9609fe59c | avg cpu time: 0.27333003382743
generator] (2018-09-18 18:53:11.820, 2018-09-18 18:53:13.467) | Daniel | c72f3629-4e9b-453f-b06d-bb95547fcdf7 | avg cpu time: 0.6650131410260026
generator] (2018-09-18 18:53:11.830, 2018-09-18 18:53:16.987) | Olga | 24cefdfe-cb1e-4b75-9d6d-5f3f8281d007 | avg cpu time: 0.7647590080392109
generator] (2018-09-18 18:53:11.863, 2018-09-18 18:53:19.111) | Daniel | 001c59a2-d38c-4f28-90ae-8e440bc64101 | avg cpu time: 0.5094041136983717
database] 2018-09-18 18:53:12.827 | Kathryn | bd8bdf8c-8836-494d-bbb7-c25920ebd9bc | putSize: 863; returnSize: 1519
database] 2018-09-18 18:53:12.840 | Sunil | c10de4d5-7959-45c5-8dfd-087f4b7aacb1 | putSize: 890; returnSize: 2231
database] 2018-09-18 18:53:12.859 | Maria | 7b50a63e-eacf-45c1-863e-c741df8009ac | putSize: 461; returnSize: 2475
database] 2018-09-18 18:53:12.880 | Sunil | ccf86d2c-83b9-40ef-9a48-94a02a1da808 | putSize: 673; returnSize: 2766
verification] 2018-09-18 18:53:13.845 | Kathryn | bd8bdf8c-8836-494d-bbb7-c25920ebd9bc | records verified: 6466
verification] 2018-09-18 18:53:13.874 | Sunil | c10de4d5-7959-45c5-8dfd-087f4b7aacb1 | records verified: 15639
verification] 2018-09-18 18:53:13.879 | Maria | 7b50a63e-eacf-45c1-863e-c741df8009ac | records verified: 8468
verification] 2018-09-18 18:53:13.897 | Sunil | ccf86d2c-83b9-40ef-9a48-94a02a1da808 | records verified: 1485
generator] (2018-09-18 18:53:14.862, 2018-09-18 18:53:16.052) | Kathryn | bd8bdf8c-8836-494d-bbb7-c25920ebd9bc | avg cpu time: 0.2232376764032
generator] (2018-09-18 18:53:14.891, 2018-09-18 18:53:21.252) | Sunil | c10de4d5-7959-45c5-8dfd-087f4b7aacb1 | avg cpu time: 0.5534759404694979
generator] (2018-09-18 18:53:14.897, 2018-09-18 18:53:20.856) | Maria | 7b50a63e-eacf-45c1-863e-c741df8009ac | avg cpu time: 0.9254998132812464

APACHE FLUME

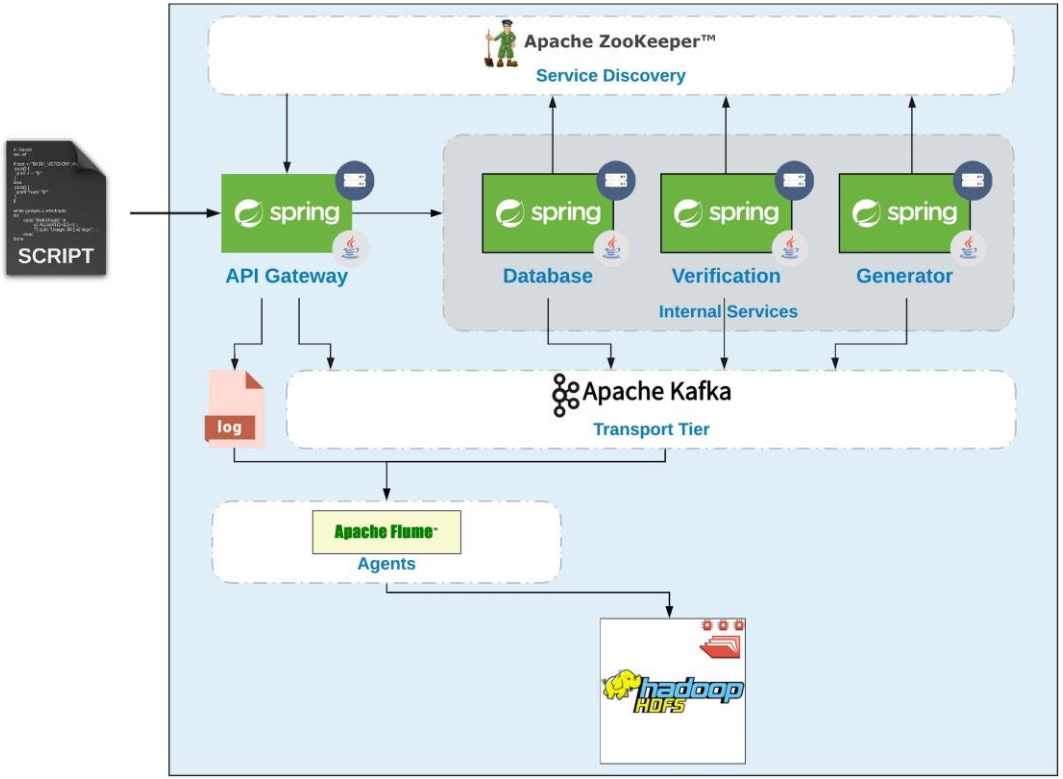
APACHE FLUME



APACHE FLUME



APPLICATION



AGENT

```
target_agent.sources = kafkaSource
target_agent.channels = memoryChannel
target_agent.sinks = hdfsSink

target_agent.sources.kafkaSource.type = org.apache.flume.source.kafka.KafkaSource
target_agent.sources.kafkaSource.zookeeperConnect = sandbox.hortonworks.com:2181
target_agent.sources.kafkaSource.topic = resource.utilization.topic
target_agent.sources.kafkaSource.channels = memoryChannel
target_agent.sources.kafkaSource.groupId = flume
target_agent.sources.kafkaSource.interceptors = i1
target_agent.sources.kafkaSource.interceptors.i1.type=timestamp
target_agent.sources.kafkaSource.consumer.timeout.ms=100

# http://flume.apache.org/FlumeUserGuide.html#memory-channel
target_agent.channels.memoryChannel.type = memory
target_agent.channels.memoryChannel.capacity = 10000
target_agent.channels.memoryChannel.transactionCapacity = 1000

## Write to HDFS
#http://flume.apache.org/FlumeUserGuide.html#hdfs-sink
target_agent.sinks.hdfsSink.type = hdfs
target_agent.sinks.hdfsSink.channel = memoryChannel
target_agent.sinks.hdfsSink.hdfs.path = hdfs://sandbox-hdp.hortonworks.com:8020/user/workshop/%{topic}
target_agent.sinks.hdfsSink.hdfs.fileType = DataStream
target_agent.sinks.hdfsSink.hdfs.writeFormat = Text
target_agent.sinks.hdfsSink.hdfs.rollSize = 104857600
target_agent.sinks.hdfsSink.hdfs.rollInterval = 0
target_agent.sinks.hdfsSink.hdfs.rollCount = 0
```

flume/flume_agent.conf

AGENT

```
target_agent.sources = kafkaSource
target_agent.channels = memoryChannel
target_agent.sinks = hdfsSink

target_agent.sources.kafkaSource.type = org.apache.flume.source.kafka.KafkaSource
target_agent.sources.kafkaSource.zookeeperConnect = sandbox.hortonworks.com:2181
target_agent.sources.kafkaSource.topic = resource.utilization.topic
target_agent.sources.kafkaSource.channels = memoryChannel
target_agent.sources.kafkaSource.groupId = flume
target_agent.sources.kafkaSource.interceptors = i1
target_agent.sources.kafkaSource.interceptors.i1.type=timestamp
target_agent.sources.kafkaSource.consumer.timeout.ms=100

# http://flume.apache.org/FlumeUserGuide.html#memory-channel
target_agent.channels.memoryChannel.type = memory
target_agent.channels.memoryChannel.capacity = 10000
target_agent.channels.memoryChannel.transactionCapacity = 1000

## Write to HDFS
#http://flume.apache.org/FlumeUserGuide.html#hdfs-sink
target_agent.sinks.hdfsSink.type = hdfs
target_agent.sinks.hdfsSink.channel = memoryChannel
target_agent.sinks.hdfsSink.hdfs.path = hdfs://sandbox-hdp.hortonworks.com:8020/user/workshop/%{topic}
target_agent.sinks.hdfsSink.hdfs.fileType = DataStream
target_agent.sinks.hdfsSink.hdfs.writeFormat = Text
target_agent.sinks.hdfsSink.hdfs.rollSize = 104857600
target_agent.sinks.hdfsSink.hdfs.rollInterval = 0
target_agent.sinks.hdfsSink.hdfs.rollCount = 0
```

flume/flume_agent.conf

AGENT

```
target_agent.sources = kafkaSource
target_agent.channels = memoryChannel
target_agent.sinks = hdfsSink

target_agent.sources.kafkaSource.type = org.apache.flume.source.kafka.KafkaSource
target_agent.sources.kafkaSource.zookeeperConnect = sandbox.hortonworks.com:2181
target_agent.sources.kafkaSource.topic = resource.utilization.topic
target_agent.sources.kafkaSource.channels = memoryChannel
target_agent.sources.kafkaSource.groupId = flume
target_agent.sources.kafkaSource.interceptors = i1
target_agent.sources.kafkaSource.interceptors.i1.type=timestamp
target_agent.sources.kafkaSource.consumer.timeout.ms=100

# http://flume.apache.org/FlumeUserGuide.html#memory-channel
target_agent.channels.memoryChannel.type = memory
target_agent.channels.memoryChannel.capacity = 10000
target_agent.channels.memoryChannel.transactionCapacity = 1000

## Write to HDFS
#http://flume.apache.org/FlumeUserGuide.html#hdfs-sink
target_agent.sinks.hdfsSink.type = hdfs
target_agent.sinks.hdfsSink.channel = memoryChannel
target_agent.sinks.hdfsSink.hdfs.path = hdfs://sandbox-hdp.hortonworks.com:8020/user/workshop/%{topic}
target_agent.sinks.hdfsSink.hdfs.fileType = DataStream
target_agent.sinks.hdfsSink.hdfs.writeFormat = Text
target_agent.sinks.hdfsSink.hdfs.rollSize = 104857600
target_agent.sinks.hdfsSink.hdfs.rollInterval = 0
target_agent.sinks.hdfsSink.hdfs.rollCount = 0
```

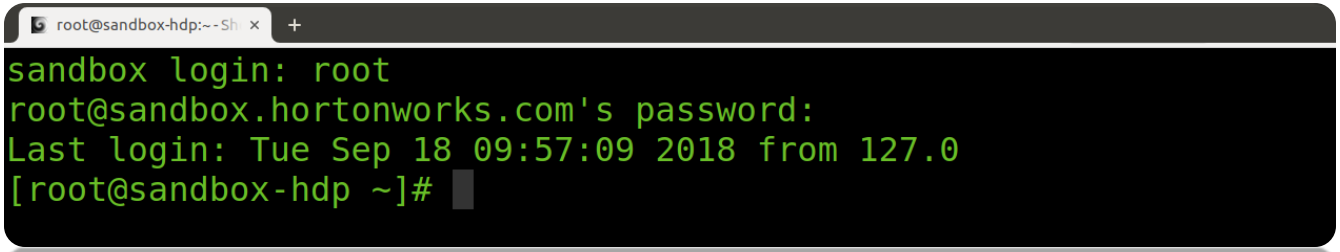
flume/flume_agent.conf

AGENT

```
target_agent.sources = kafkaSource
target_agent.channels = memoryChannel
target_agent.sinks = hdfsSink

target_agent.sources.kafkaSource.type = org.apache.flume.source.kafka.KafkaSource
target_agent.sources.kafkaSource.zookeeperConnect = sandbox.hortonworks.com:2181
target_agent.sources.kafkaSource.topic = resource.utilization.topic
target_agent.sources.kafkaSource.channels = memoryChannel
target_agent.sources.kafkaSource.groupId = flume
target_agent.sources.kafkaSource.interceptors = i1
target_agent.sources.kafkaSource.interceptors.i1.type=timestamp
target_agent.sources.kafkaSource.consumer.timeout.ms=100

# http://flume.apache.org/FlumeUserGuide.html#memory-channel
target_agent.channels.memoryChannel.type = memory
target_agent.channels.memoryChannel.capacity = 10000
target_agent.channels.memoryChannel.transactionCapacity = 1000

## Write to HDFS
#http://flume.apache.org/FlumeUserGuide.html#hdfs-sink
target_agent.sinks.hdfsSink.type = hdfs
target_agent.sinks.hdfsSink.channel = memoryChannel
target_agent.sinks.hdfsSink.hdfs.path = hdfs://sandbox-hdp.hortonworks.com:8020/user/workshop/%{topic}
target_agent.sinks.hdfsSink.hdfs.fileType = DataStream
target_agent.sinks.hdfsSink.hdfs.writeFormat = Text
target_agent.sinks.hdfsSink.hdfs.rollSize = 104857600
target_agent.sinks.hdfsSink.hdfs.rollInterval = 0
target_agent.sinks.hdfsSink.hdfs.rollCount = 0
```

flume/flume_agent.conf

LET'S PLAY



- Login

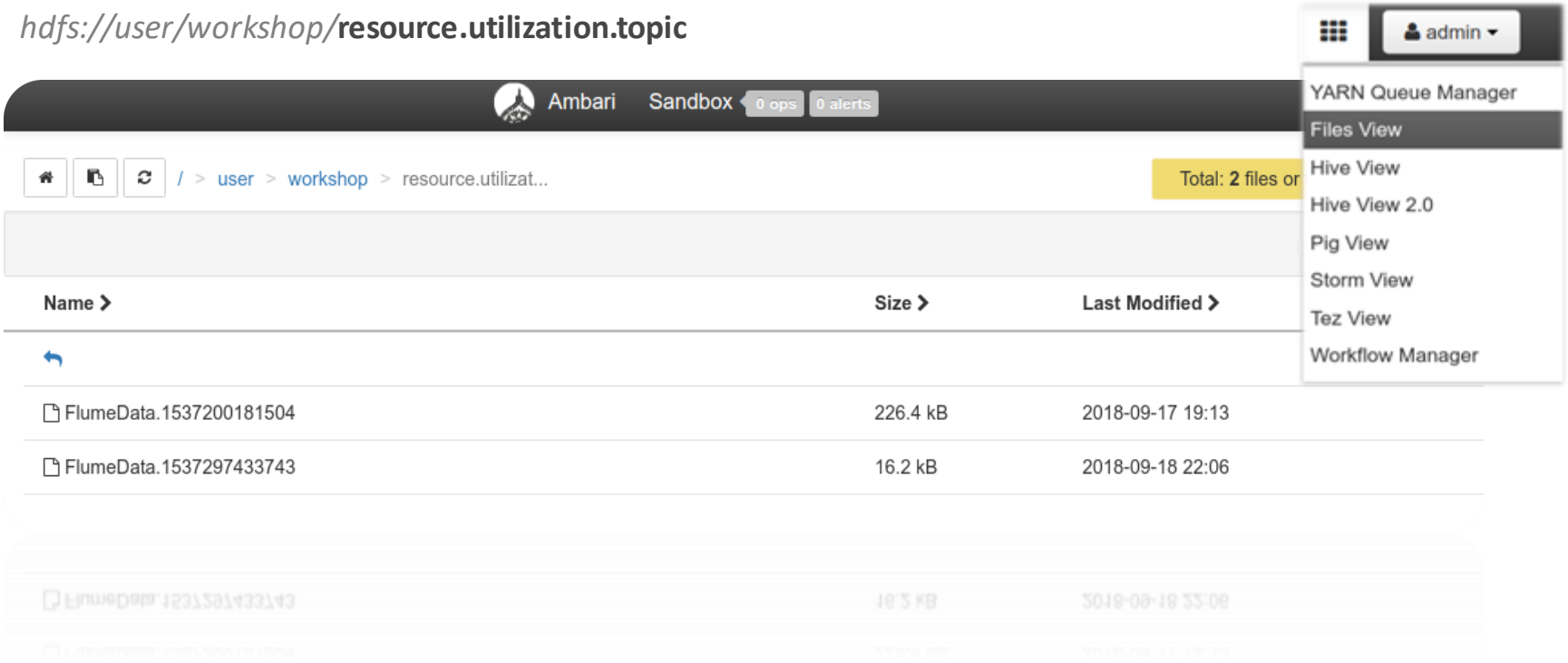
  address: http://127.0.0.1:4200/ , creds: **root / workshop**

- Run

```
# cd workshop/flume
# /usr/hdp/current/flume-server/bin/flume-ng agent
    --conf /usr/hdp/current/flume-server/conf
    --conf-file flume_agent.conf
    --name target_agent
    -Dflume.root.logger=INFO,console
```

LET'S PLAY

*hdfs://user/workshop/***resource.utilization.topic**

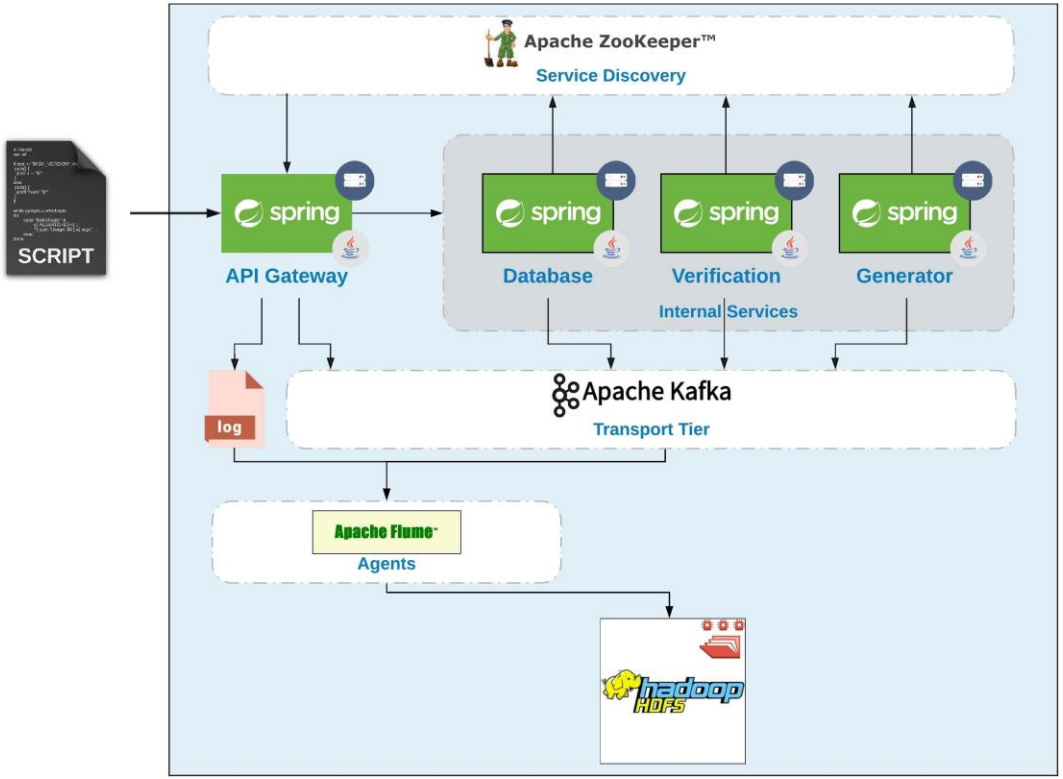


http://127.0.0.1:8080

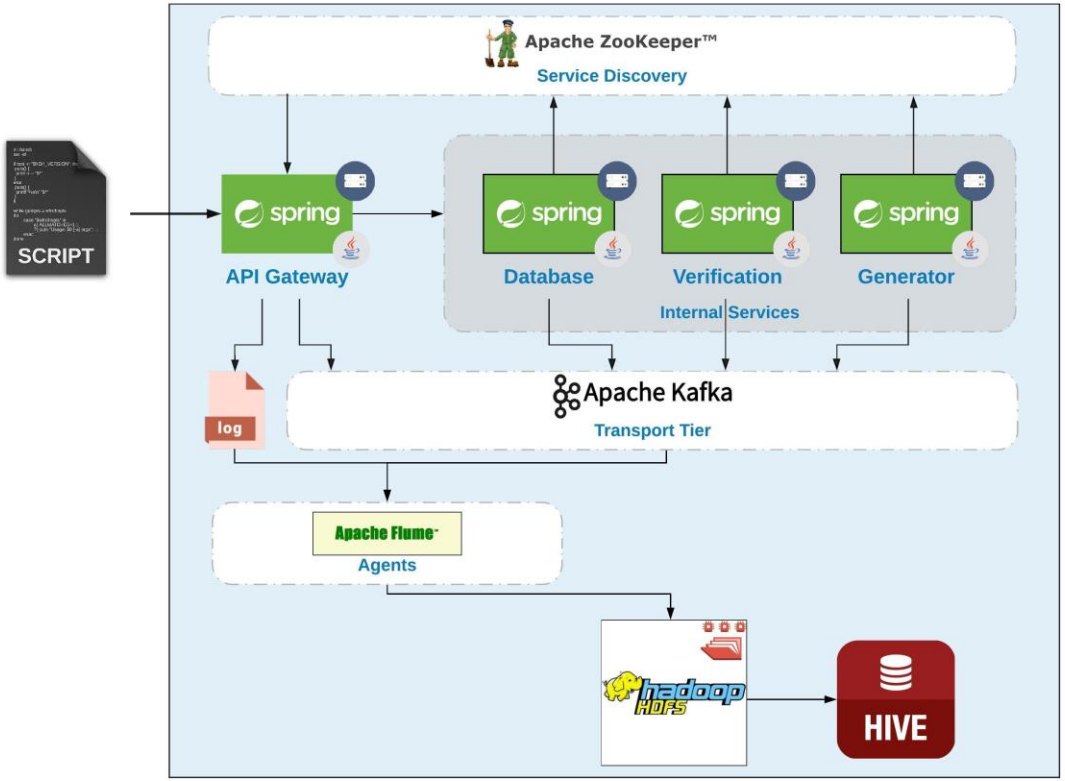
http://127.0.0.1:8080/#/main/view/FILES/auto_files_instance

MAP-REDUCE

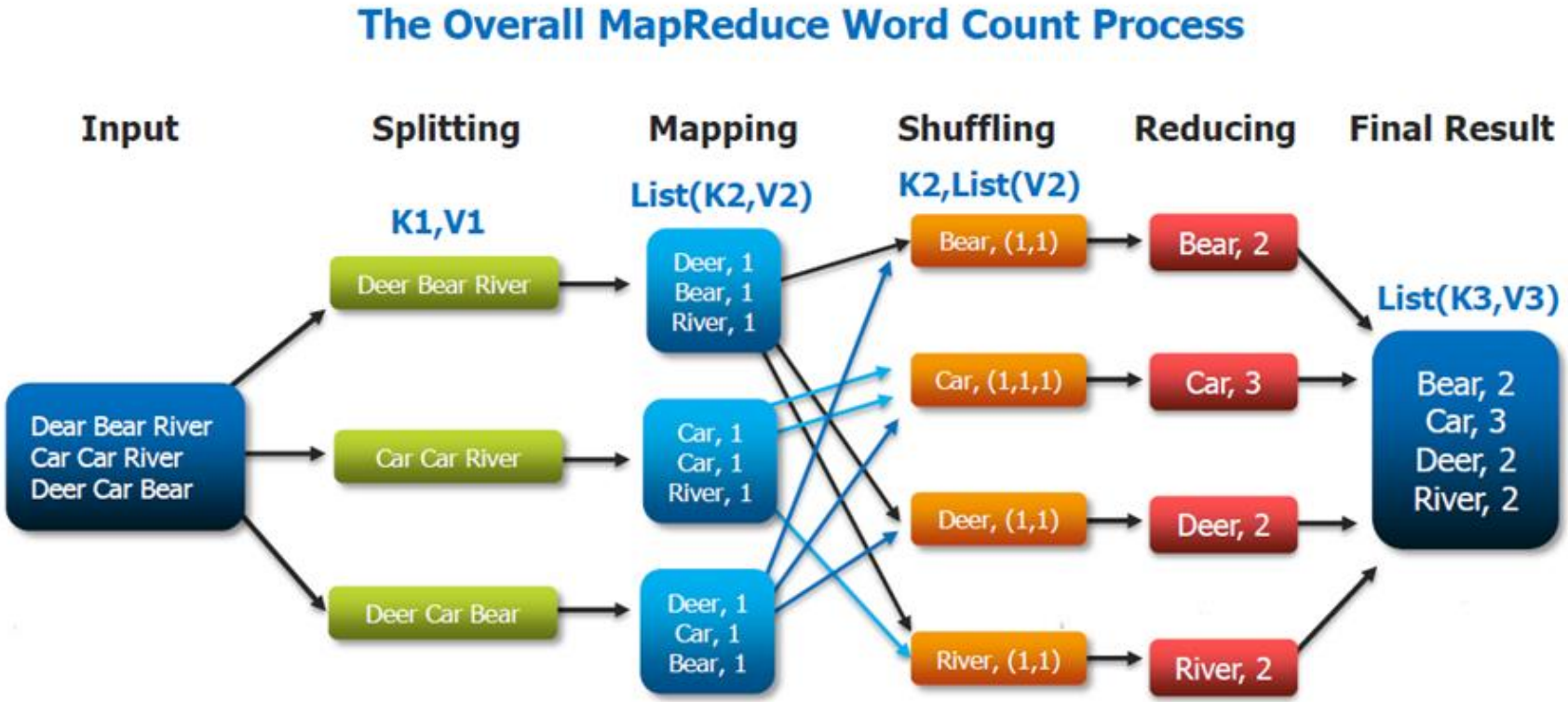
APPLICATION



APPLICATION



MAP-REDUCE


The Overall MapReduce Word Count Process
Input
Splitting
Mapping
Shuffling
Reducing
Final Result
K1,V1
List(K2,V2)
K2,List(V2)
List(K3,V3)
Dear Bear River
Car Car River
Deer Car Bear
Deer Bear River
Car Car River
Deer Car Bear
Deer, 1
Bear, 1
River, 1
Car, 1
Car, 1
River, 1
Deer, 1
Car, 1
Bear, 1
Bear, (1,1)
Car, (1,1,1)
Deer, (1,1)
River, (1,1)
Bear, 2
Car, 3
Deer, 2
River, 2
Bear, 2
Car, 3
Deer, 2
River, 2

MESSAGES

database-service

```
"[database] {ts} | {userId} | {workflowId} | putSize: {szInKb}; returnSize: {szIbKb}"
  -> resource.utilization.topic
```

verification-service

```
"[verification] {ts} | {userId} | {workflowId} | records verified: {number}"
  -> resource.utilization.topic
```

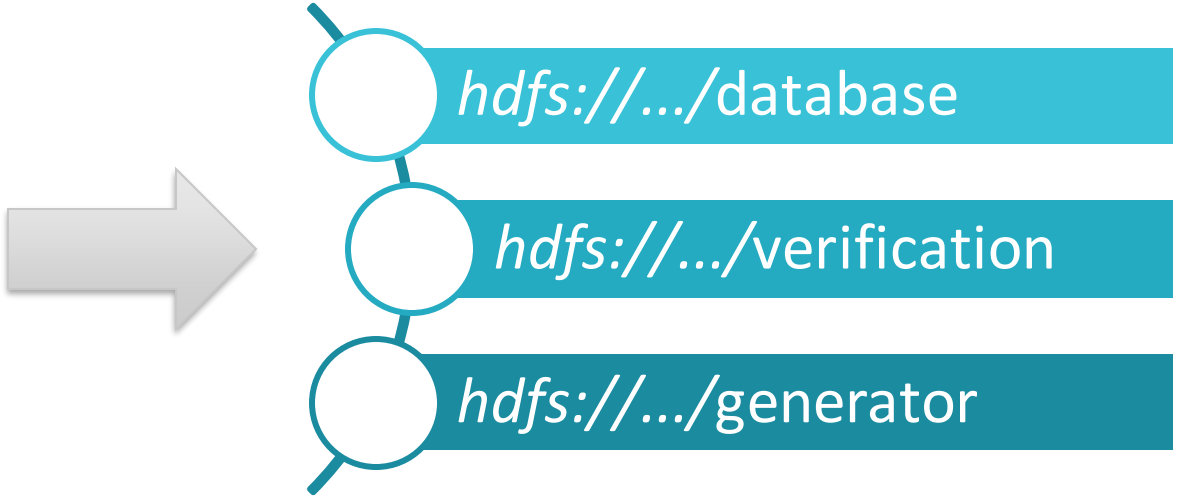
generator-service

```
"[generator] ({ts1}, {ts2}) | {userId} | {workflowId} | avg cpu time: {avgCpuTime}"
  -> resource.utilization.topic
```

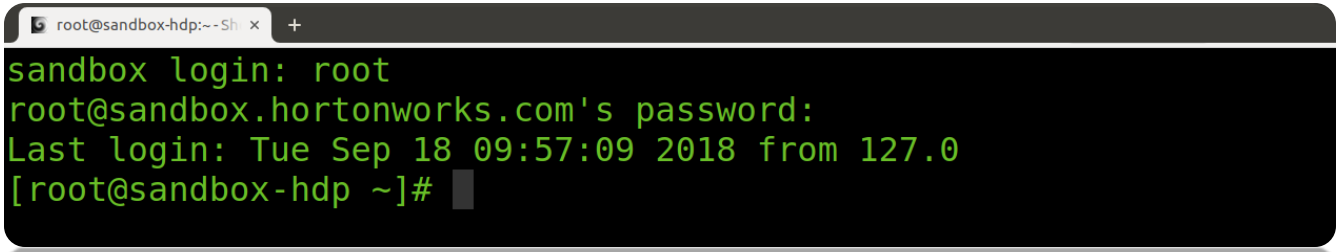
workflowId = UUID.*randomUUID*(), ts = yyyy-MM-dd HH:mm:ss.SSS

MAP-REDUCE

*hdfs://.../***resource.utilization.topic**



LET'S PLAY



- Login

  address: http://127.0.0.1:4200/ , creds: **root / workshop**

- Run

```
# cd workshop/map-reduce
# hadoop jar file-structurer-1.0-SNAPSHOT-jar-with-dependencies.jar
        com.epam.bdcc.workshop.structurer.driver.StructurerDriver
        /user/workshop/resource.utilization.topic
        /user/workshop/hive_data
```

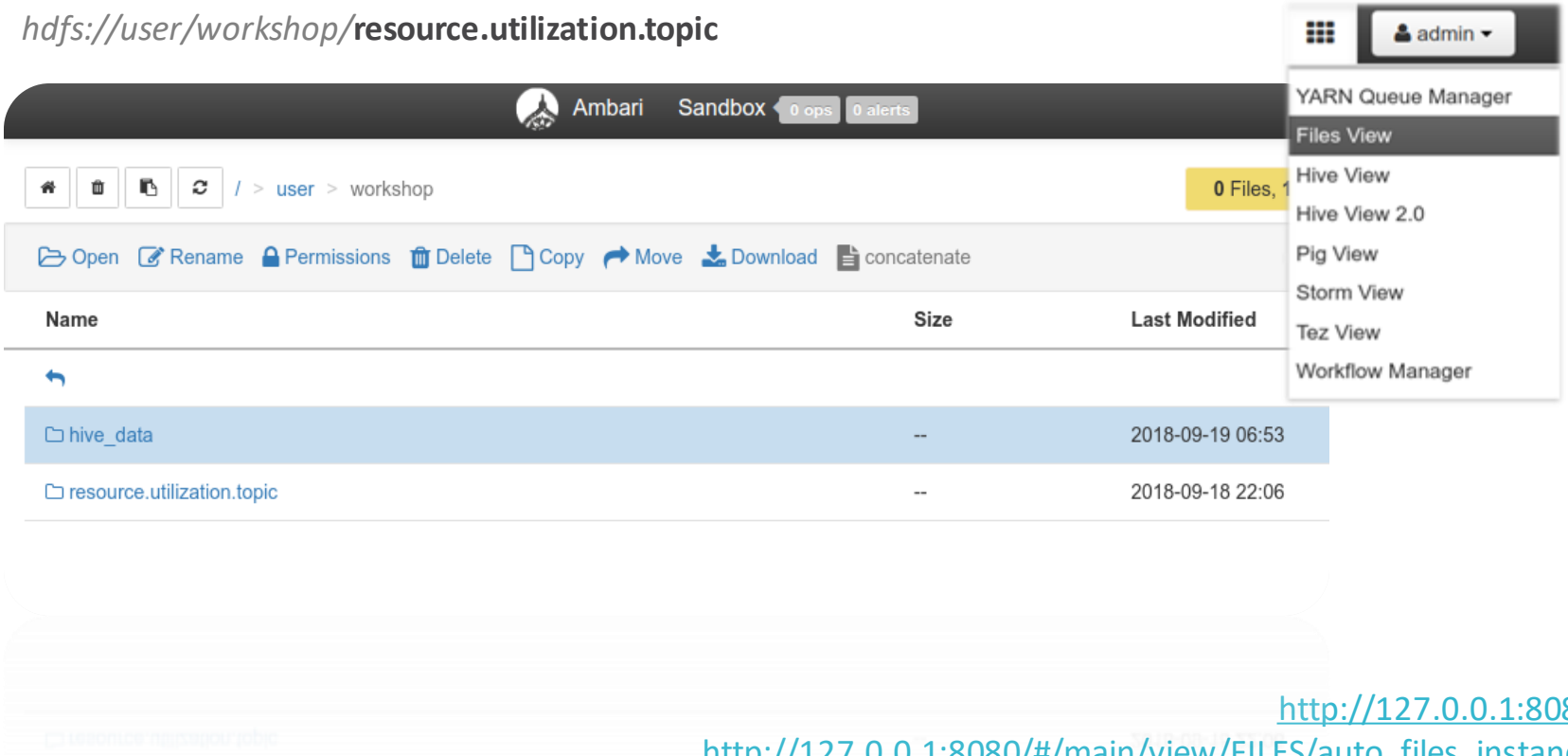
LET'S PLAY

```
18/09/19 04:09:59 INFO client.RMProxy: Connecting to ResourceManager at sandbox-hdp.hortonworks.com/172.17.0.2:8032
18/09/19 04:09:59 INFO client.AHSProxy: Connecting to Application History server at sandbox-hdp.hortonworks.com/172.17.0.2:10200
18/09/19 04:10:00 INFO input.FileInputFormat: Total input paths to process : 2
18/09/19 04:10:00 INFO mapreduce.JobSubmitter: number of splits:2
18/09/19 04:10:00 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1537328875388_0002
18/09/19 04:10:00 INFO impl.YarnClientImpl: Submitted application application_1537328875388_0002
18/09/19 04:10:00 INFO mapreduce.Job: The url to track the job: http://sandbox-hdp.hortonworks.com:8088/proxy/application_1537328875388_0002/
18/09/19 04:10:00 INFO mapreduce.Job: Running job: job_1537328875388_0002
18/09/19 04:10:07 INFO mapreduce.Job: Job job_1537328875388_0002 running in uber mode : false
18/09/19 04:10:07 INFO mapreduce.Job:  map 0% reduce 0%
18/09/19 04:10:14 INFO mapreduce.Job:  map 100% reduce 0%
18/09/19 04:10:20 INFO mapreduce.Job:  map 100% reduce 100%
18/09/19 04:10:21 INFO mapreduce.Job: Job job_1537328875388_0002 completed successfully
18/09/19 04:10:21 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=365961
                FILE: Number of bytes written=1192109
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=248739
                HDFS: Number of bytes written=165699
                HDFS: Number of read operations=9
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=8872
                Total time spent by all reduces in occupied slots (ms)=3901
                Total time spent by all map tasks (ms)=8872
                Total time spent by all reduce tasks (ms)=3901
                Total vcore-milliseconds taken by all map tasks=8872
                Total vcore-milliseconds taken by all reduce tasks=3901
                Total megabyte-milliseconds taken by all map tasks=2218000
                Total megabyte-milliseconds taken by all reduce tasks=975250
        Map-Reduce Framework
                Map input records=2181
                Map output records=2181
                Map output bytes=361549
                Map output materialized bytes=365967
                Input split bytes=338
                Combine input records=0
                Combine output records=0
                Reduce input groups=2181
                Reduce shuffle bytes=365967
                Reduce input records=2181
                Reduce output records=0
                Spilled Records=4362
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=736
                CPU time spent (ms)=2600
                Physical memory (bytes) snapshot=545206272
                Virtual memory (bytes) snapshot=6405017600
                Total committed heap usage (bytes)=274726912
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=248401
        File Output Format Counters
                Bytes Written=0
```


```
18/09/19 04:10:00 INFO mapreduce.Job: Running job: job_1537328875388_0002
18/09/19 04:10:07 INFO mapreduce.Job: Job job 1537328875388_0002 running in uber mode : false
18/09/19 04:10:07 INFO mapreduce.Job:  map 0% reduce 0%
18/09/19 04:10:14 INFO mapreduce.Job:  map 100% reduce 0%
18/09/19 04:10:20 INFO mapreduce.Job:  map 100% reduce 100%
18/09/19 04:10:21 INFO mapreduce.Job: Job job 1537328875388_0002 completed successfully
18/09/19 04:10:21 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=365961
                FILE: Number of bytes written=1192109
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=248739
                HDFS: Number of bytes written=165699
                HDFS: Number of read operations=9
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=8872
                Total time spent by all reduces in occupied slots (ms)=3901
                Total time spent by all map tasks (ms)=8872
                Total time spent by all reduce tasks (ms)=3901
                Total vcore-milliseconds taken by all map tasks=8872
                Total vcore-milliseconds taken by all reduce tasks=3901
                Total megabyte-milliseconds taken by all map tasks=2218000
                Total megabyte-milliseconds taken by all reduce tasks=975250
```

LET'S PLAY

*hdfs://user/workshop/***resource.utilization.topic**

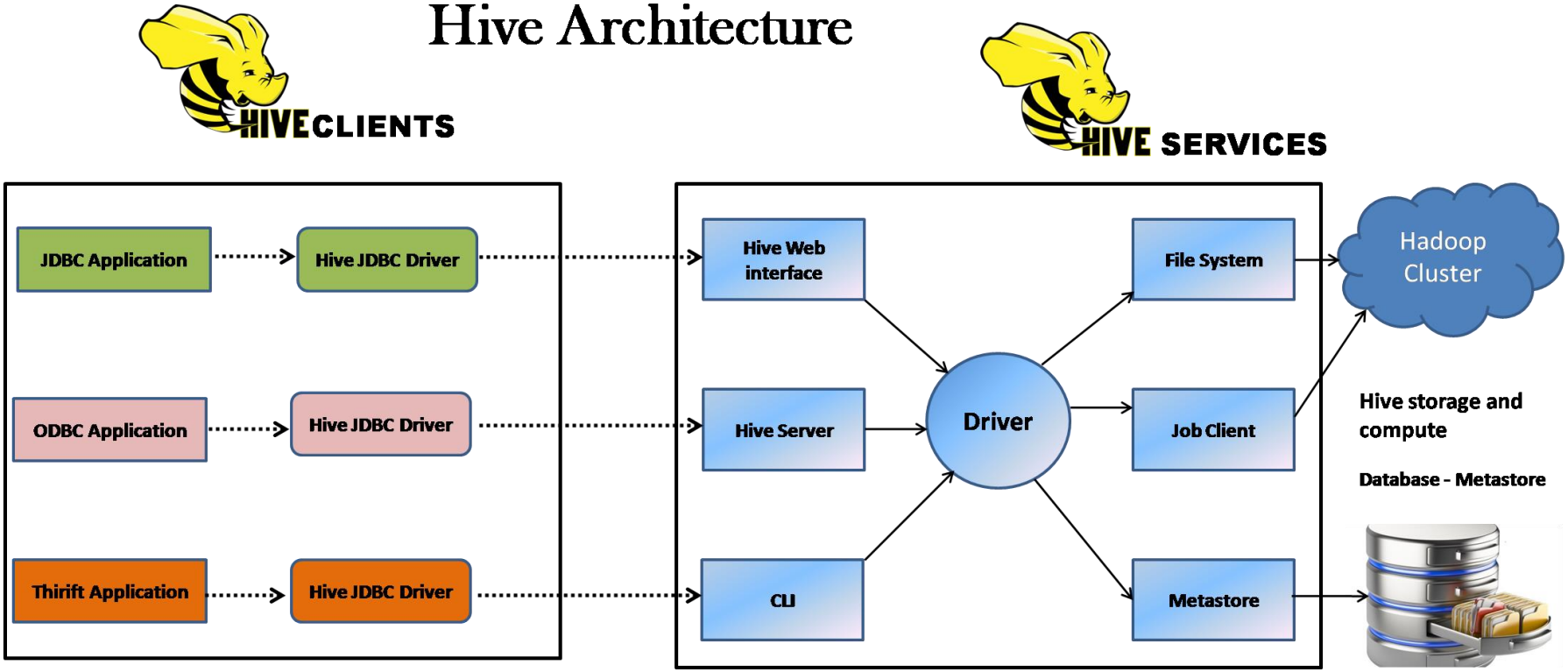


http://127.0.0.1:8080

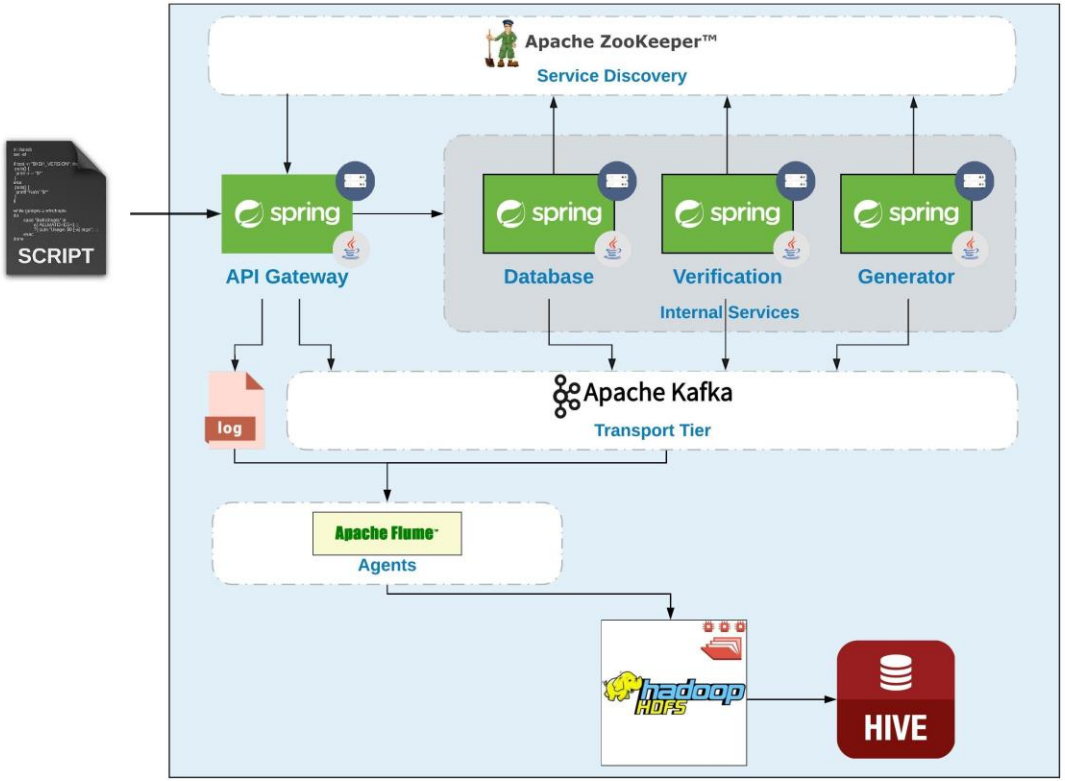
http://127.0.0.1:8080/#/main/view/FILES/auto_files_instance

HIVE

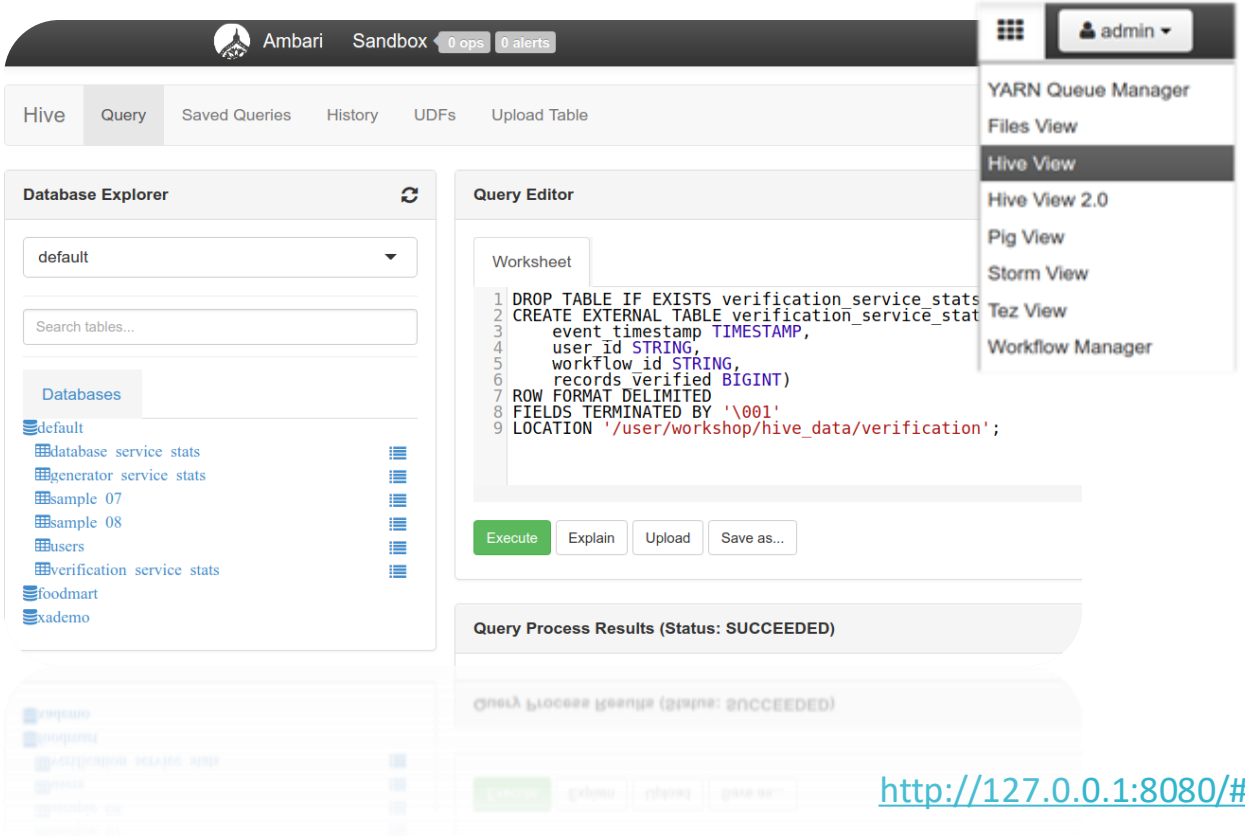
Hive Architecture



APPLICATION



LET'S PLAY



http://127.0.0.1:8080

http://127.0.0.1:8080/#/main/view/HIVE/auto_hive_instance

LET'S PLAY

```
DROP TABLE IF EXISTS verification_service_stats;
CREATE EXTERNAL TABLE verification_service_stats (
    event_timestamp TIMESTAMP,
    user_id STRING,
    workflow_id STRING,
    records_verified BIGINT)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '\001'
LOCATION '/user/workshop/hive_data/verification';
```

hive/create_verification_table.txt

http://127.0.0.1:8080
http://127.0.0.1:8080/#/main/view/HIVE/auto_hive_instance

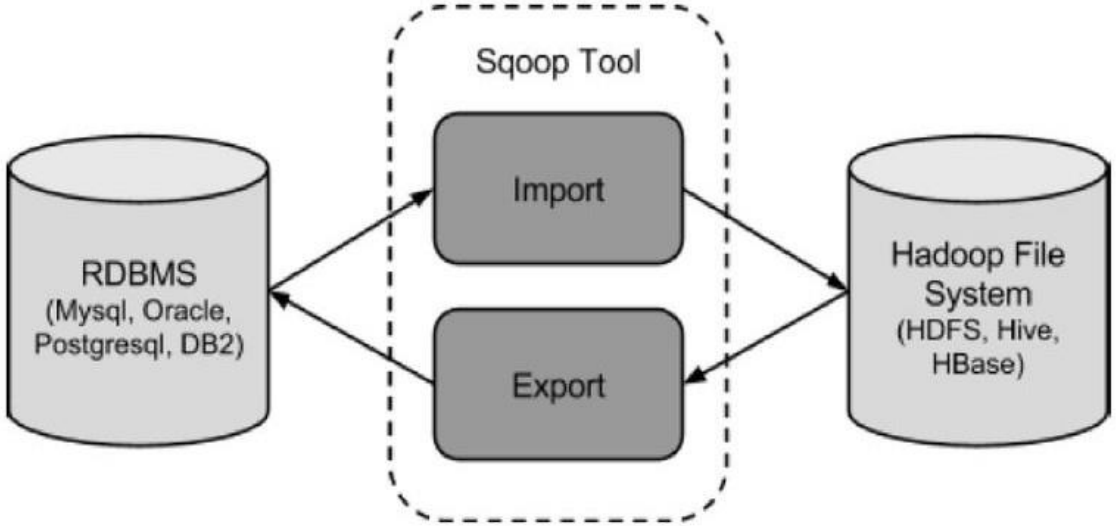
LET'S PLAY

```
with
  tbl_with_summed_sizes as (
    select db.user_id, db.workflow_id,
           db.put_size + db.return_size as summed_sizes,
           generator.avg_cpu_time
    from database_service_stats db inner join generator_service_stats generator
    on (db.workflow_id = generator.workflow_id)
 ),
 ranked_tbl_with_summed_sizes as (
   select dense_rank() over (order by twss.summed_sizes desc) as ranked_summed_sizes,
   twss.* from tbl_with_summed_sizes twss
 )

select * from ranked_tbl_with_summed_sizes where ranked_summed_sizes = 1;
```

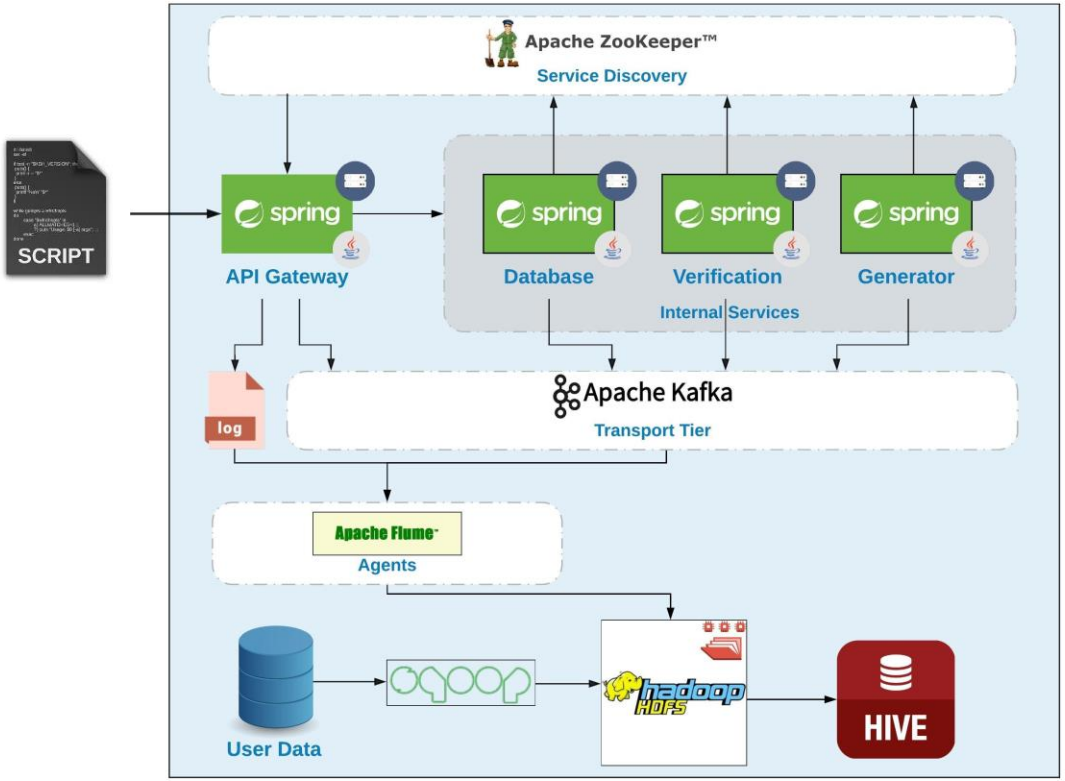
hive/top_database_disk_consumption_user.txt

SQOOP

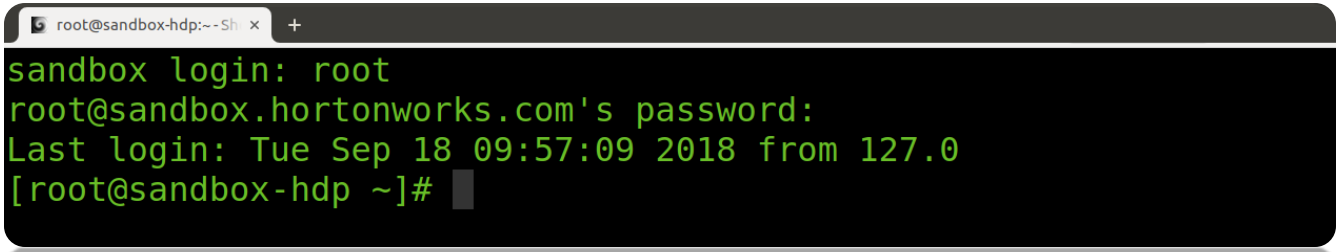
SQOOP


Sqoop Tool
Import
Export
RDBMS
(Mysql, Oracle,
Postgresql, DB2)
Hadoop File
System
(HDFS, Hive,
HBase)

APPLICATION



LET'S PLAY



- Login

  address: http://127.0.0.1:4200/ , creds: **root / workshop**

- Go to sqoop directory

  ```
  # cd workshop/sqoop
  ```

LET'S PLAY

```
# mysql -p -u root                                    (password = hadoop)

create database organizations;
use organizations;
create table users(
 id int primary key,
 organization varchar(50),
 user varchar(100)
);
insert into users values(1, 'EPAM', 'Dmitrii');
insert into users values(2, 'EPAM', 'Olga');
insert into users values(3, 'T-Systems', 'Jennifer');
insert into users values(4, 'T-Systems', 'Kathryn');
insert into users values(5, 'T-Systems', 'Sunil');
insert into users values(6, 'Luxoft', 'Maria');
insert into users values(7, 'Luxoft', 'Alexander');
exit;
```

sqoop/instructions_v1.txt

LET'S PLAY

```
# set hive.warehouse.subdir.inherit.perms = false;

# /usr/hdp/current/sqoop-server/bin/sqoop import
  --connect jdbc:mysql://localhost:3306/organizations
  --table users --username root -P
  --hive-import --create-hive-table --hive-table users
  --driver com.mysql.jdbc.Driver
```

(password = **hadoop**)

sqoop/instructions_v1.txt

LET'S PLAY

```
18/09/19 04:46:41 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6.2.6.4.0-91
Enter password:
18/09/19 04:46:45 INFO tool.BaseSqoopTool: Using Hive-specific delimiters for output. You can override
18/09/19 04:46:45 INFO tool.BaseSqoopTool: delimiters with --fields-terminated-by, etc.
18/09/19 04:46:45 WARN sqoop.ConnFactory: Parameter --driver is set to an explicit driver however appropriate connection manager is not being set (via -
--connection-manager). Sqoop is going to fall back to org.apache.sqoop.manager.GenericJdbcManager. Please specify explicitly which connection manager sho
uld be used next time.
18/09/19 04:46:45 INFO manager.SqlManager: Using default fetchSize of 1000
18/09/19 04:46:45 INFO tool.CodeGenTool: Beginning code generation
18/09/19 04:46:46 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM users AS t WHERE 1=0
18/09/19 04:46:46 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM users AS t WHERE 1=0
18/09/19 04:46:46 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/hdp/2.6.4.0-91/hadoop-mapreduce
Note: /tmp/sqoop-root/compile/4e3268fe7be518ea94463f8248f3bfb3/users.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
18/09/19 04:46:47 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-root/compile/4e3268fe7be518ea94463f8248f3bfb3/users.jar
18/09/19 04:46:47 INFO mapreduce.ImportJobBase: Beginning import of users
18/09/19 04:46:48 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM users AS t WHERE 1=0
18/09/19 04:46:49 INFO client.RMProxy: Connecting to ResourceManager at sandbox-hdp.hortonworks.com/172.17.0.2:8032
18/09/19 04:46:49 INFO client.AHSProxy: Connecting to Application History server at sandbox-hdp.hortonworks.com/172.17.0.2:10200
18/09/19 04:46:54 INFO db.DBInputFormat: Using read commited transaction isolation
18/09/19 04:46:54 INFO db.DataDrivenDBInputFormat: BoundingValsQuery: SELECT MIN(id), MAX(id) FROM users
18/09/19 04:46:54 INFO db.IntegerSplitter: Split size: 1; Num splits: 4 from: 1 to: 7
18/09/19 04:46:54 INFO mapreduce.JobSubmitter: number of splits:4
18/09/19 04:46:55 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1537328875388_0004
18/09/19 04:46:56 INFO impl.YarnClientImpl: Submitted application application_1537328875388_0004
18/09/19 04:46:56 INFO mapreduce.Job: The url to track the job: http://sandbox-hdp.hortonworks.com:8088/proxy/application_1537328875388_0004/
18/09/19 04:46:56 INFO mapreduce.Job: Running job: job_1537328875388_0004
18/09/19 04:47:03 INFO mapreduce.Job: Job job_1537328875388_0004 running in uber mode : false
18/09/19 04:47:03 INFO mapreduce.Job:  map 0% reduce 0%
18/09/19 04:47:11 INFO mapreduce.Job:  map 25% reduce 0%
18/09/19 04:47:13 INFO mapreduce.Job:  map 50% reduce 0%
18/09/19 04:47:14 INFO mapreduce.Job:  map 100% reduce 0%
18/09/19 04:47:15 INFO mapreduce.Job: Job job_1537328875388_0004 completed successfully
18/09/19 04:47:16 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=682672
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=377
                HDFS: Number of bytes written=120
                HDFS: Number of read operations=16
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=8
        Job Counters
                Launched map tasks=4
                Other local map tasks=4
                Total time spent by all maps in occupied slots (ms)=25856
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=25856
                Total vcore-milliseconds taken by all map tasks=25856
                Total megabyte-milliseconds taken by all map tasks=6464000
```

Query Editor

users sample

```
1 SELECT * FROM users LIMIT 100;
```

Execute | Explain | Upload | Save as...

Query Process Results (Status: SUCCEEDED)

Logs | Results

Filter columns...

users.id	users.organization	users.user
1	EPAM	Dmitrii
2	EPAM	Olga
3	T-Systems	Jennifer
4	T-Systems	Kathryn
5	T-Systems	Sunil
6	Luxoft	Maria
7	Luxoft	Alexander

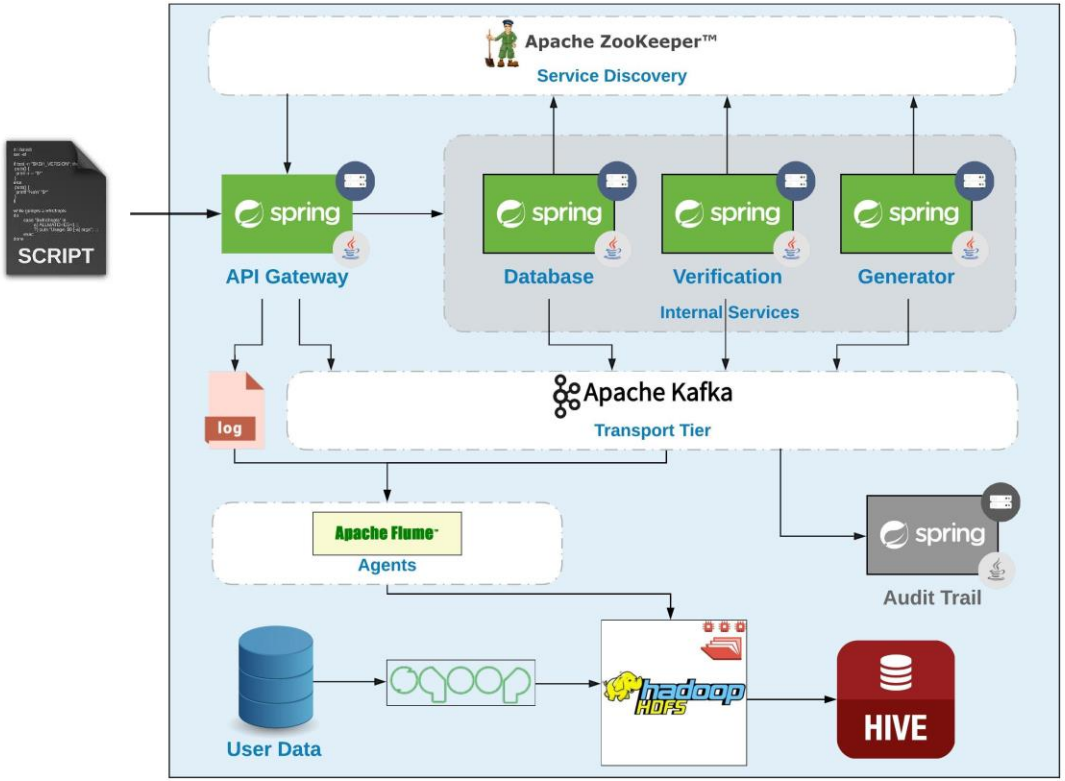
LET'S PLAY

```
with
  tbl_with_summed_sizes as (
    select usrs.organization, db.user_id, db.workflow_id,
           db.put_size + db.return_size as user_summed_sizes,
           generator.avg_cpu_time
    from database_service_stats db
    inner join generator_service_stats generator on (db.workflow_id = generator.workflow_id)
    inner join users usrs on (db.user_id = usrs.`user`)
  ),
  summed_by_organizations as (
    select organization, sum(user_summed_sizes) as org_summed_sizes
    from tbl_with_summed_sizes
    group by organization
  )

select dense_rank() over (order by sbo.org_summed_sizes desc) as ranked_org_summed_sizes,
sbo.* from summed_by_organizations sbo
```

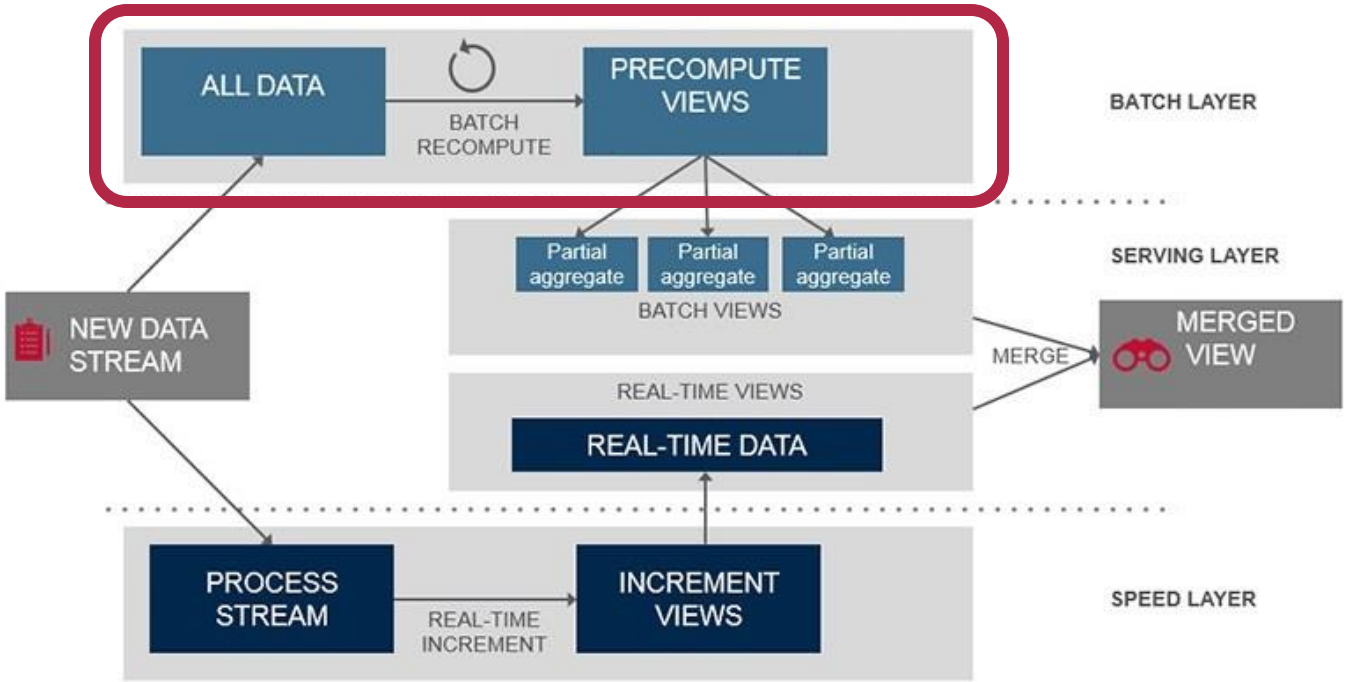
hive/top_database_disk_consumption_organization.txt

YOUR TURN

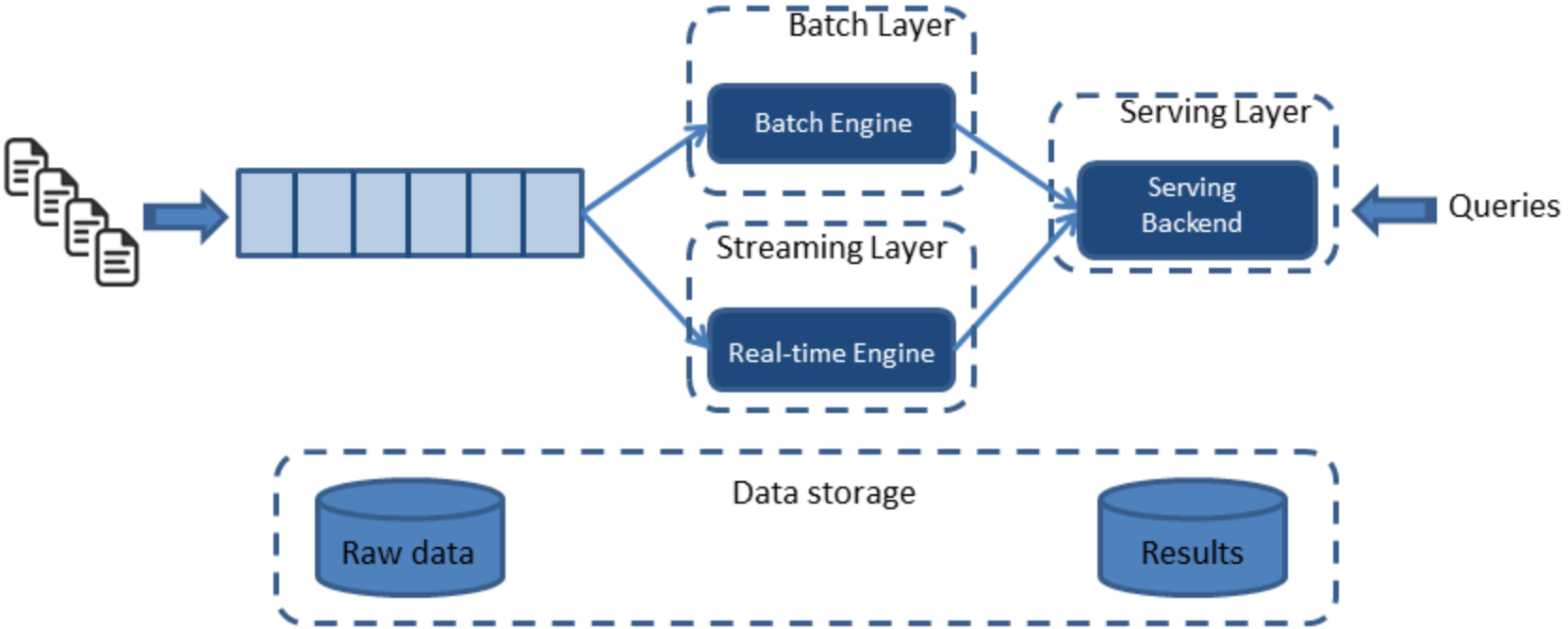
REALTIME LAYER? YOUR TURN ;)



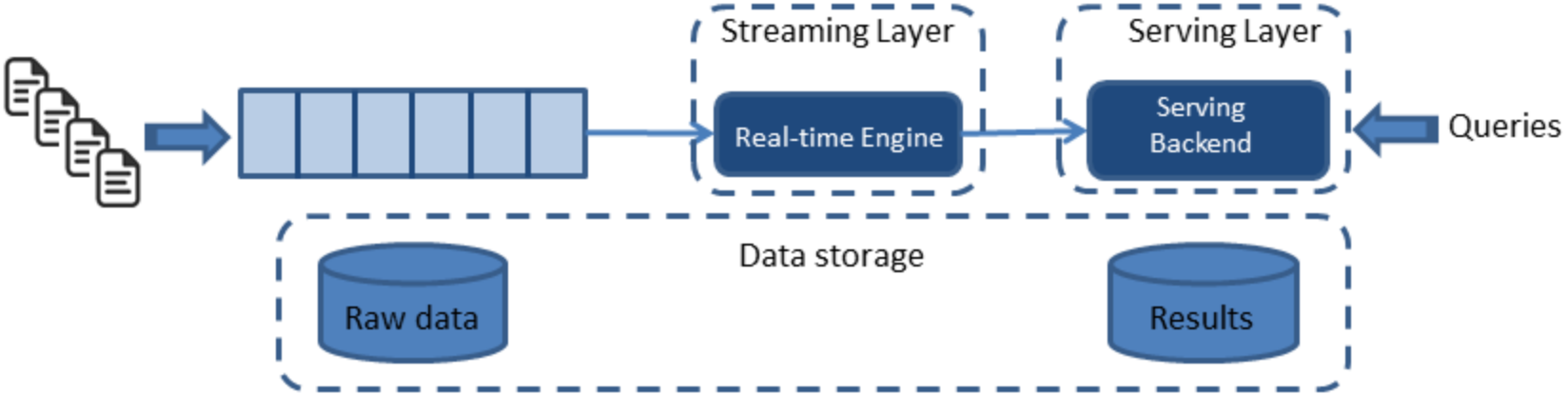
BIG DATA APPROACH: LAMBDA ARCHITECTURE



BIG DATA APPROACH: LAMBDA ARCHITECTURE



BIG DATA APPROACH: KAPPA ARCHITECTURE




THANK YOU!