

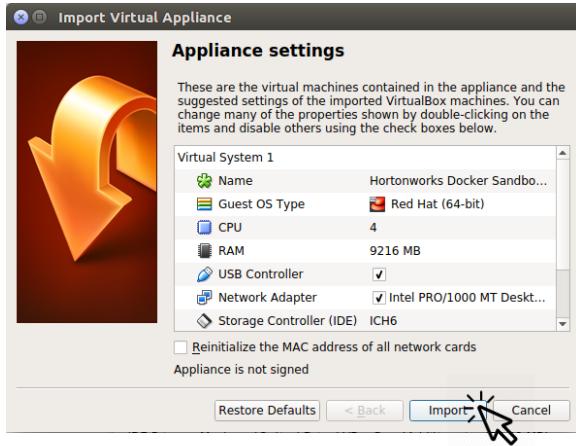
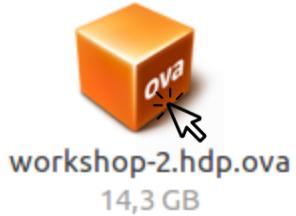


BDCC WORKSHOP

BIG DATA: THE FIRST STEP

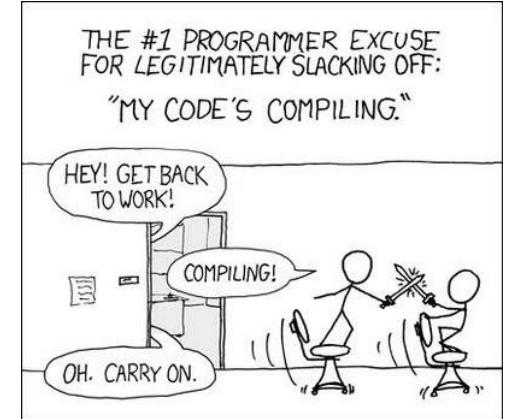
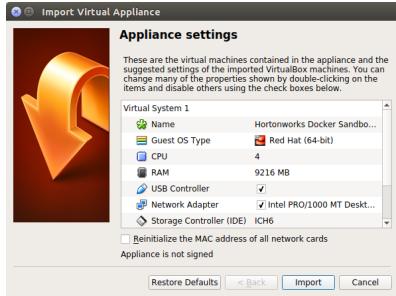
September 2018

BEFORE START



Open Import Start

BEFORE START



Open > Import > Start

...



BDCC WORKSHOP

BIG DATA: THE FIRST STEP

September 2018

AGENDA

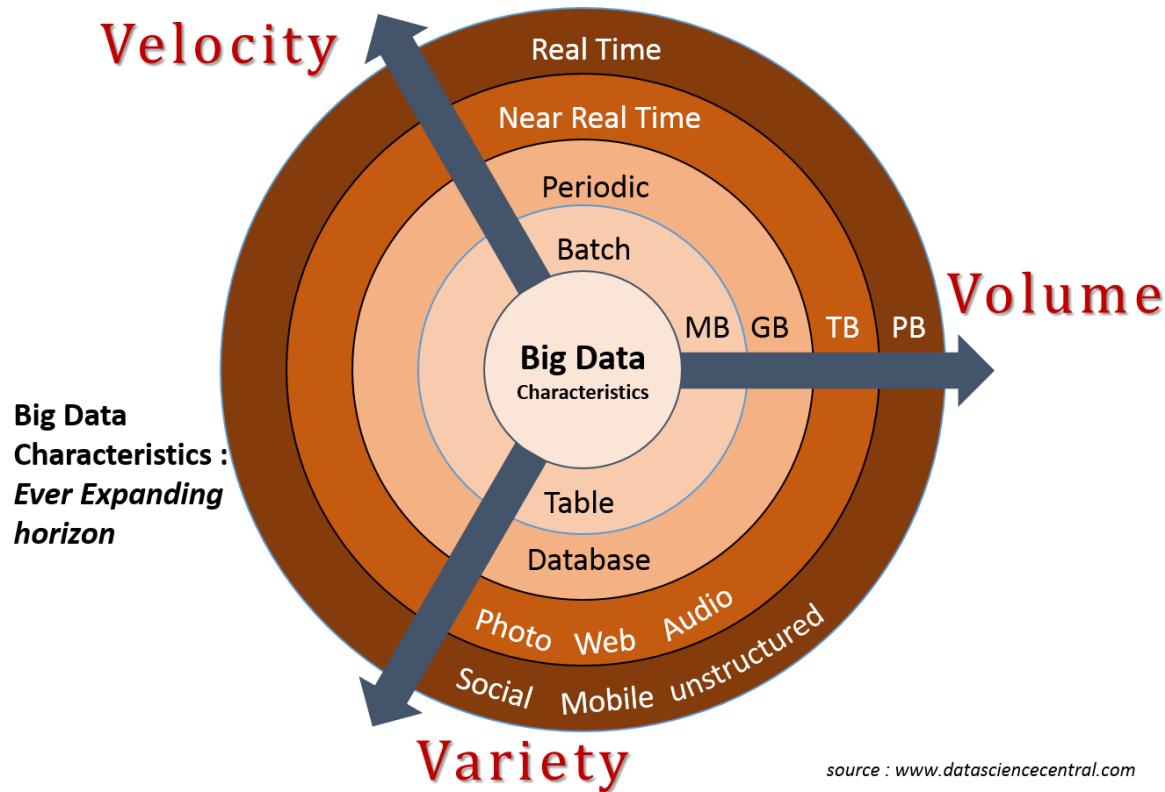
1 Big Data Overview

2 Case Study

3 Practice

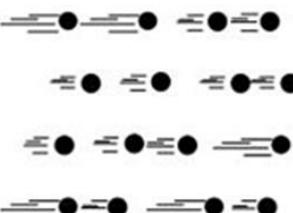
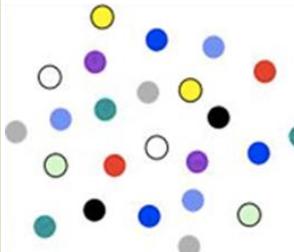
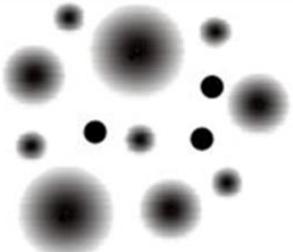
BIG DATA OVERVIEW

PREREQUISITES #1 (3V)



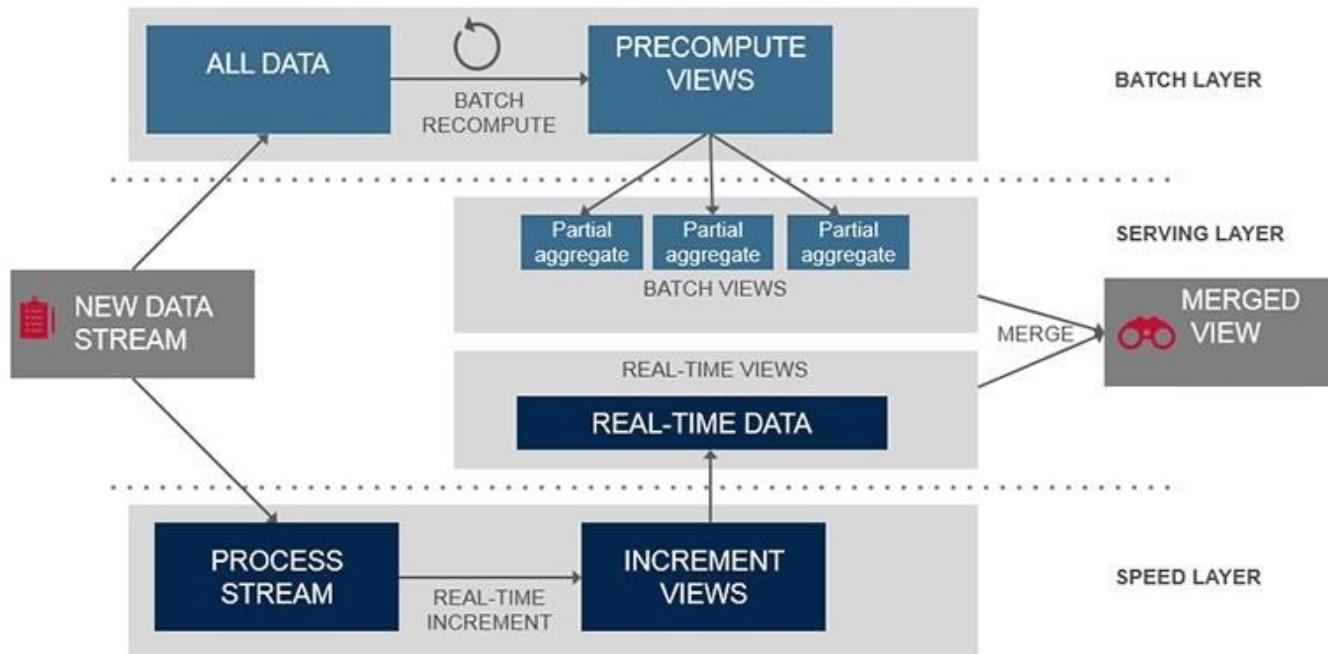
source : www.datasciencecentral.com

PREREQUISITES #2 (5V)

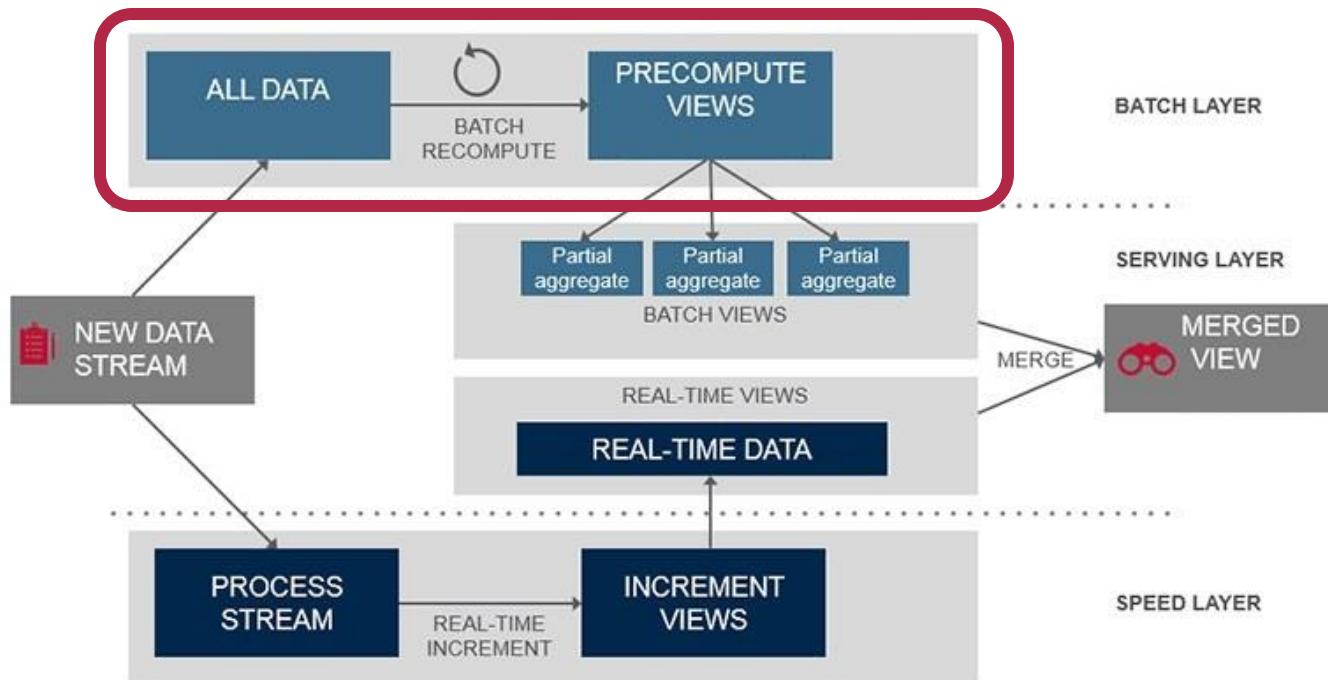
Volume	Velocity	Variety	Veracity	Value
 <p>Data at Rest Terabytes to Exabytes of existing data to process</p>	 <p>Data in Motion Streaming data, requiring milliseconds to seconds to respond</p>	 <p>Data in Many Forms Structured, unstructured, text, multimedia,...</p>	 <p>Data in Doubt Uncertainty due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

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.

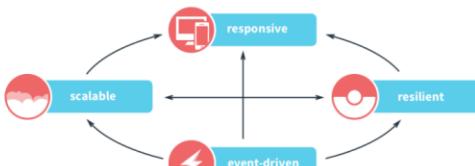
SOLUTION PRESENTATION: OVERVIEW

The solution is inspired by:

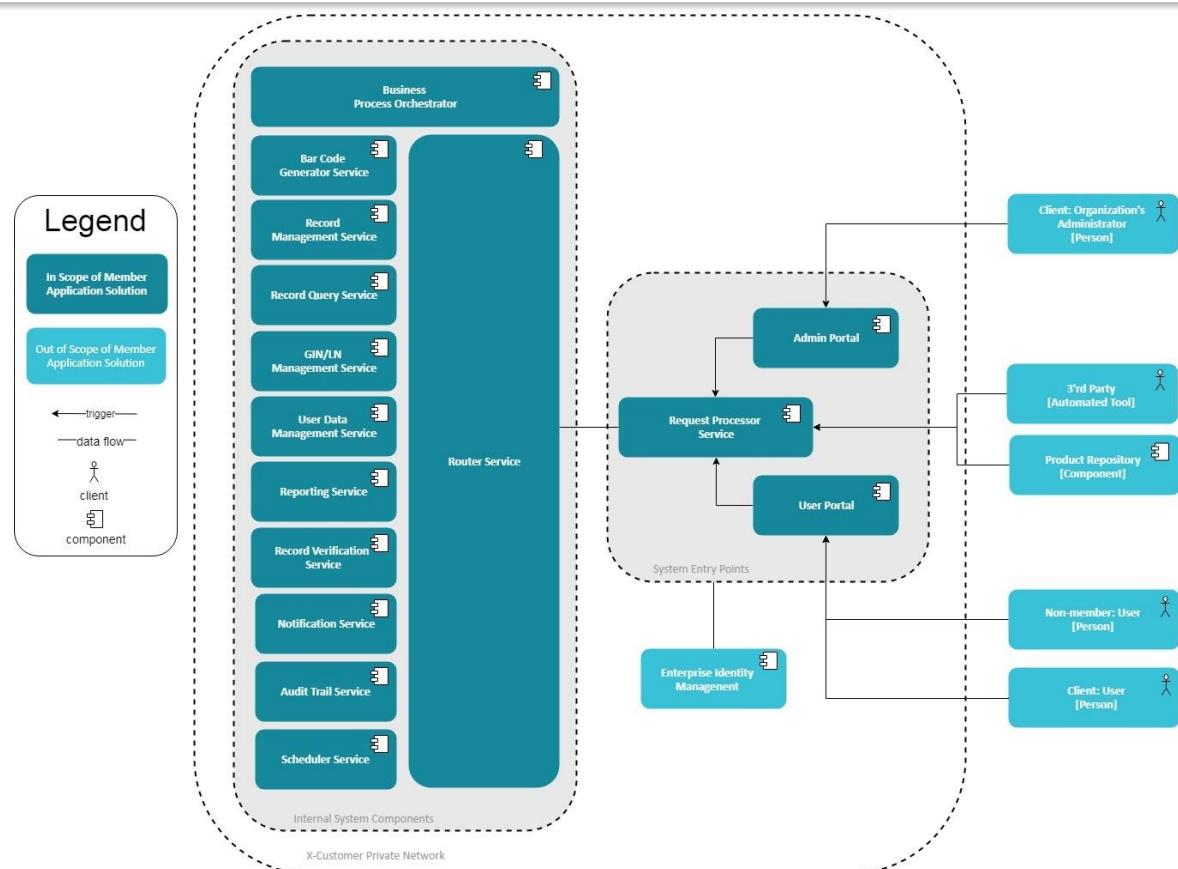
- SOA



- Reactive Architecture



- Decoupling Reads and Writes

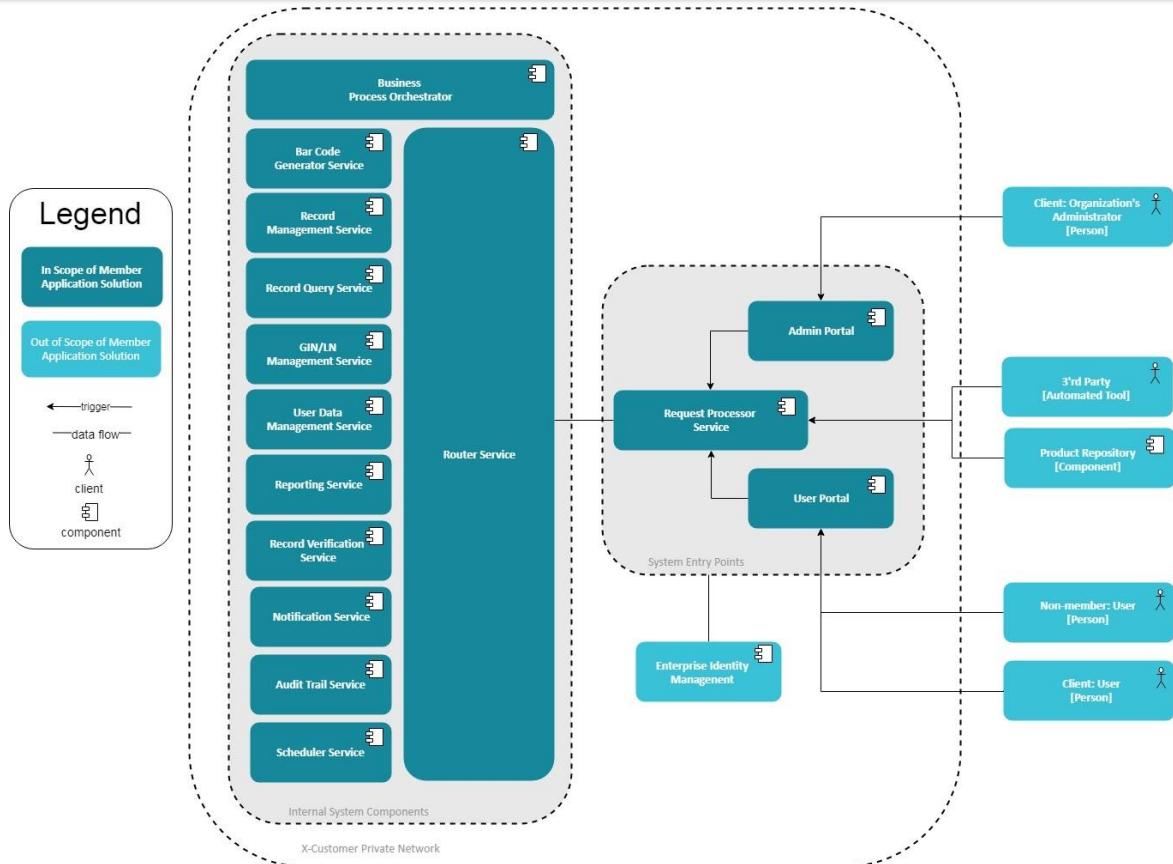


SOLUTION PRESENTATION: RATIONALE

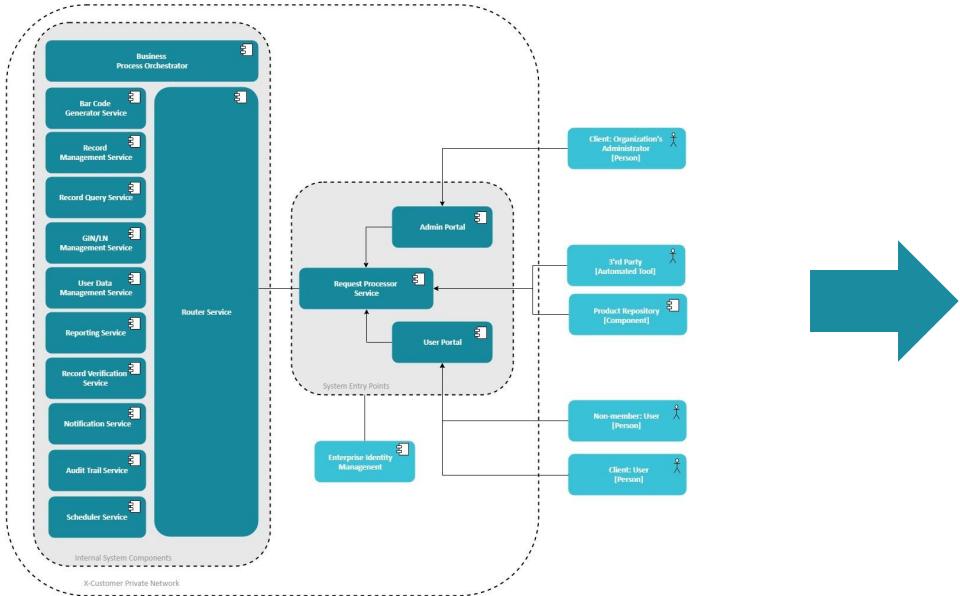


BENEFITS

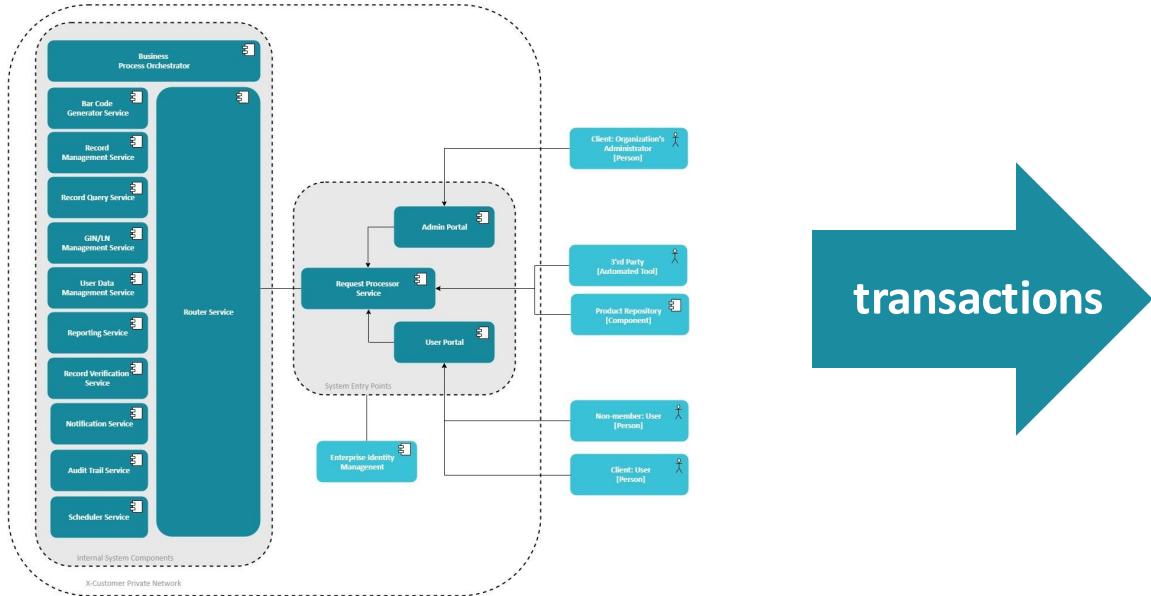
- Scalability
- High-availability
- Security
- Performance
- Fault-tolerance
- Flexibility



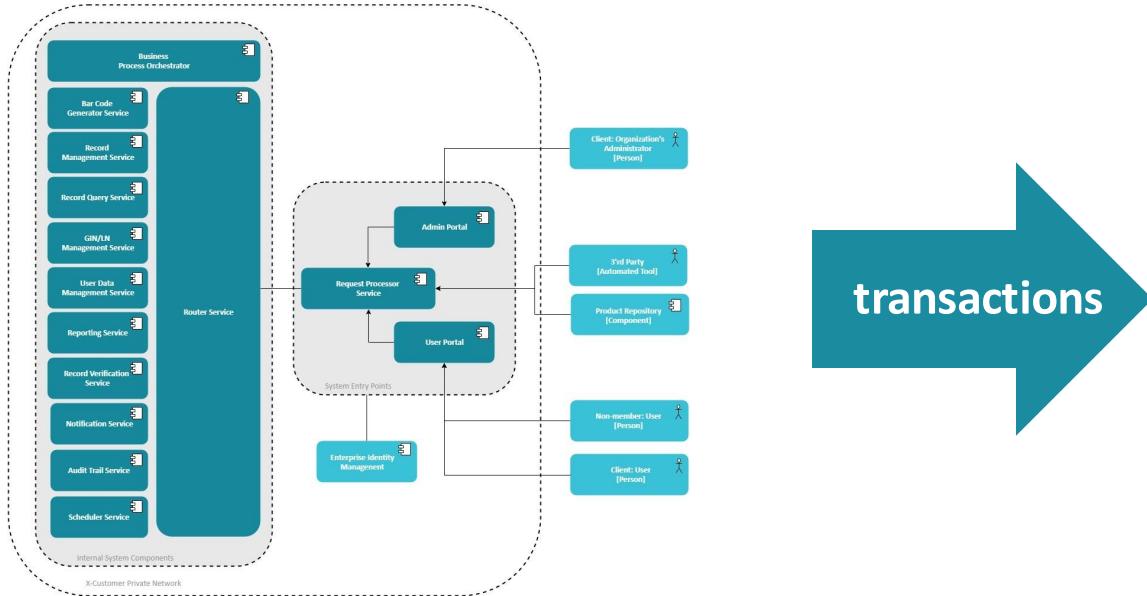
SOLUTION PRESENTATION: FOOTPRINT



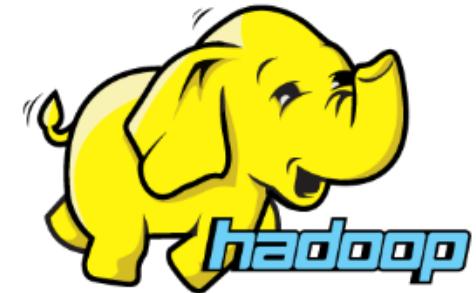
SOLUTION PRESENTATION: FOOTPRINT



SOLUTION PRESENTATION: FOOTPRINT



transactions



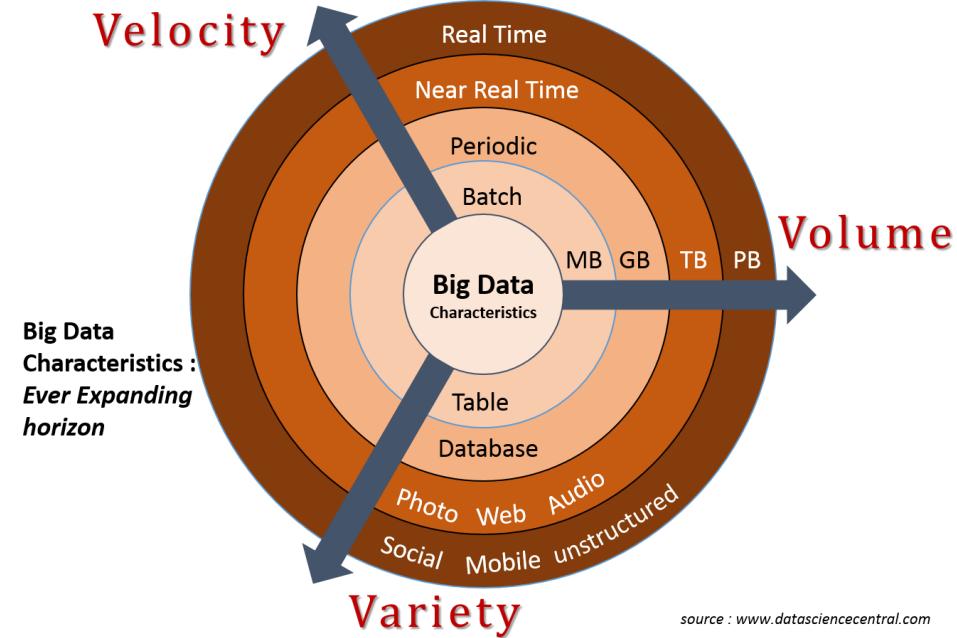
CODE GOES TO DATA

HDFS... AND HERE BIG DATA BEGINS...



SCOPE

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

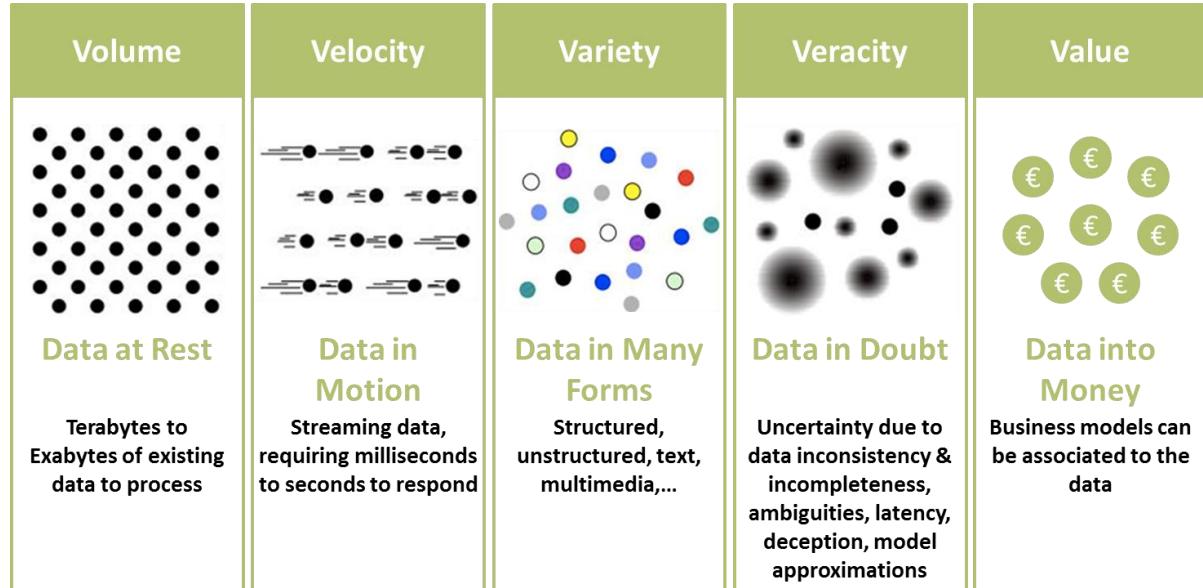


HDFS... AND HERE BIG DATA BEGINS...



SCOPE

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



Adapted by a post of Michael Walker on 28 November 2012

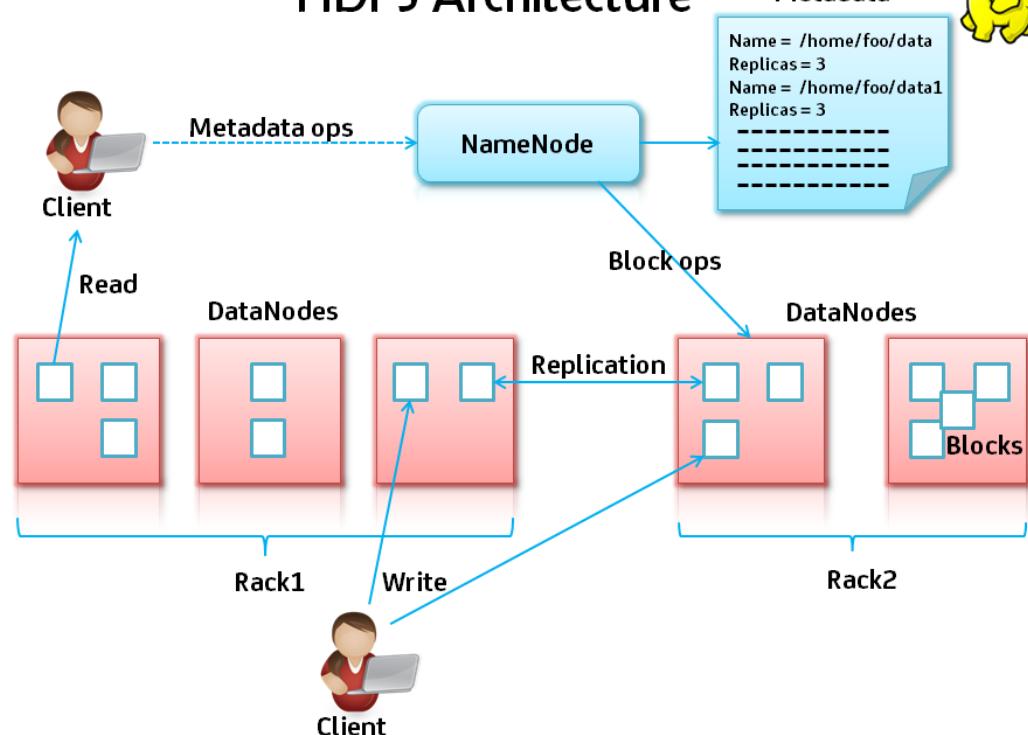
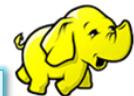
HDFS... AND HERE BIG DATA BEGINS...



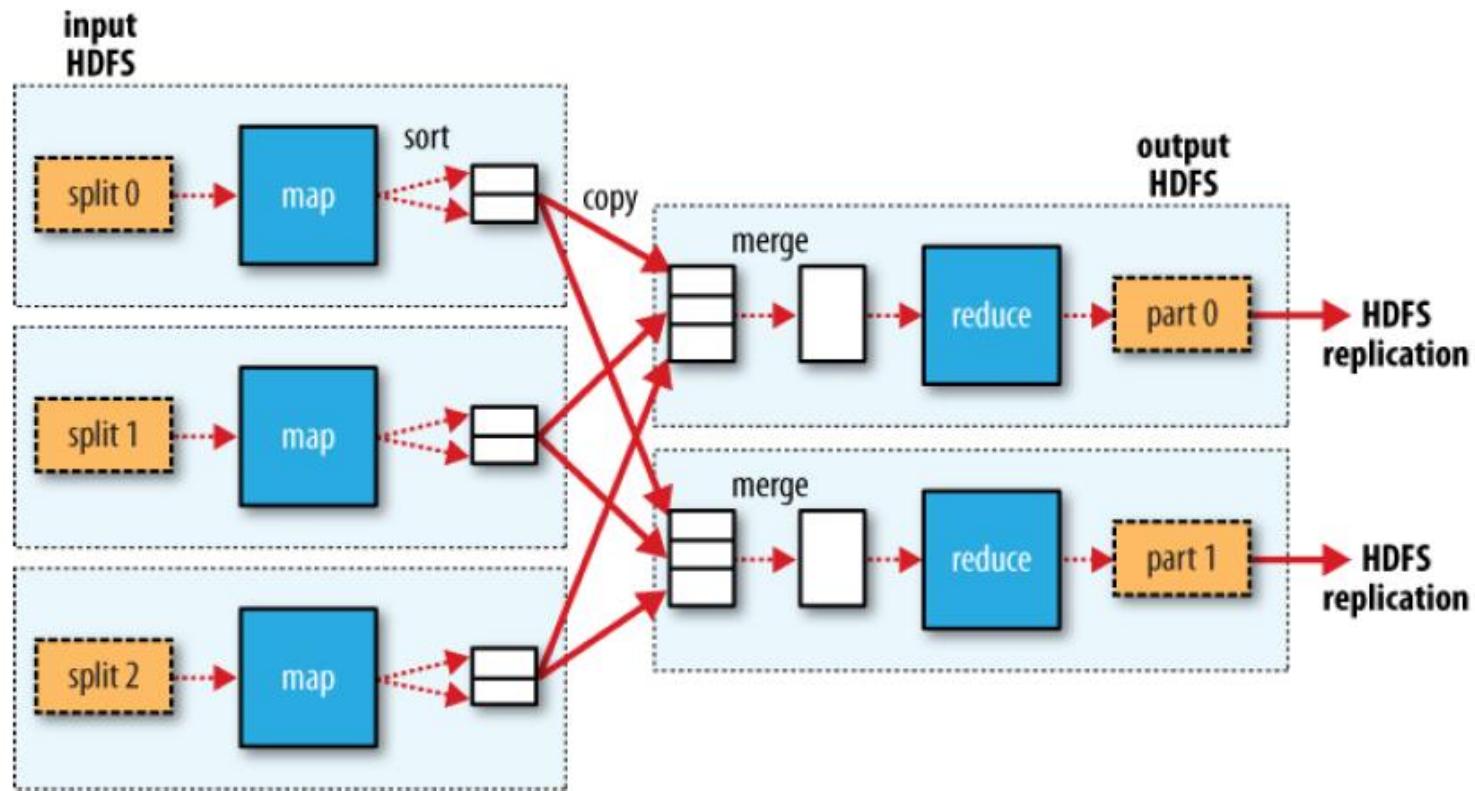
PRINCIPLES

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

HDFS Architecture

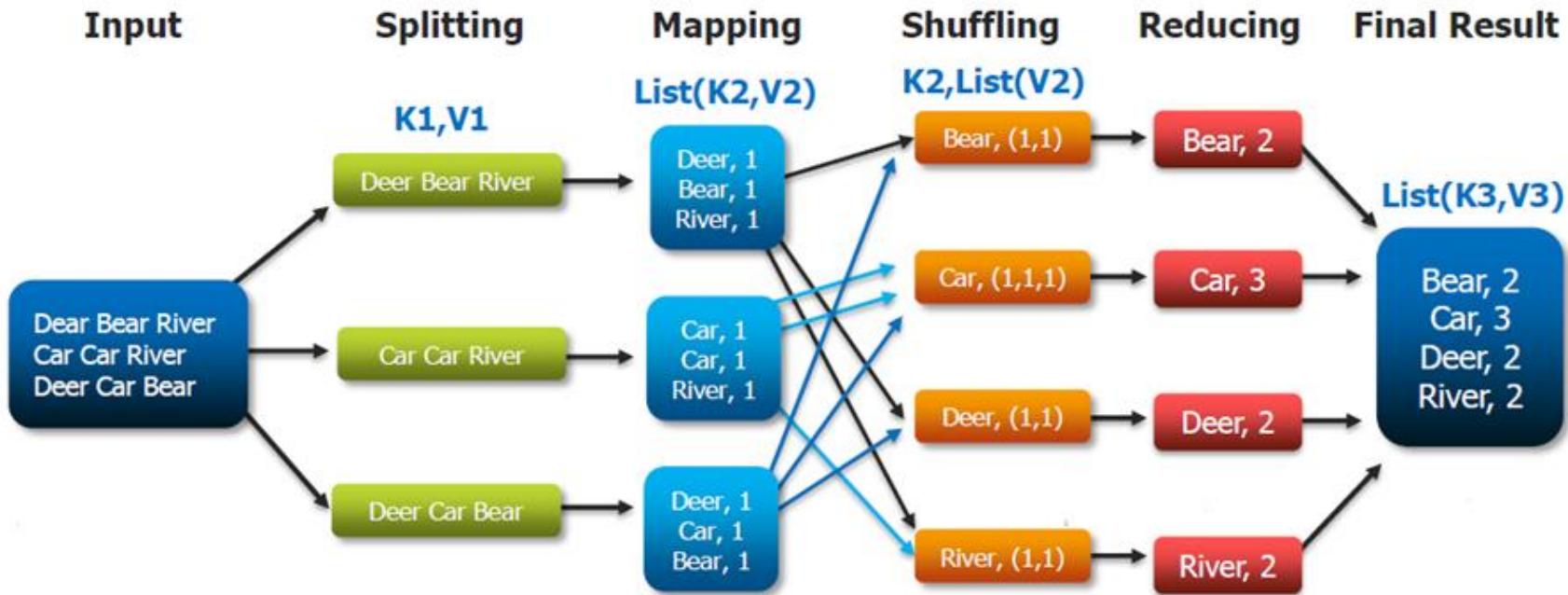


MAP-REDUCE



MAP-REDUCE

The Overall MapReduce Word Count Process



BIG DATA LANDSCAPE

BIG DATA LANDSCAPE

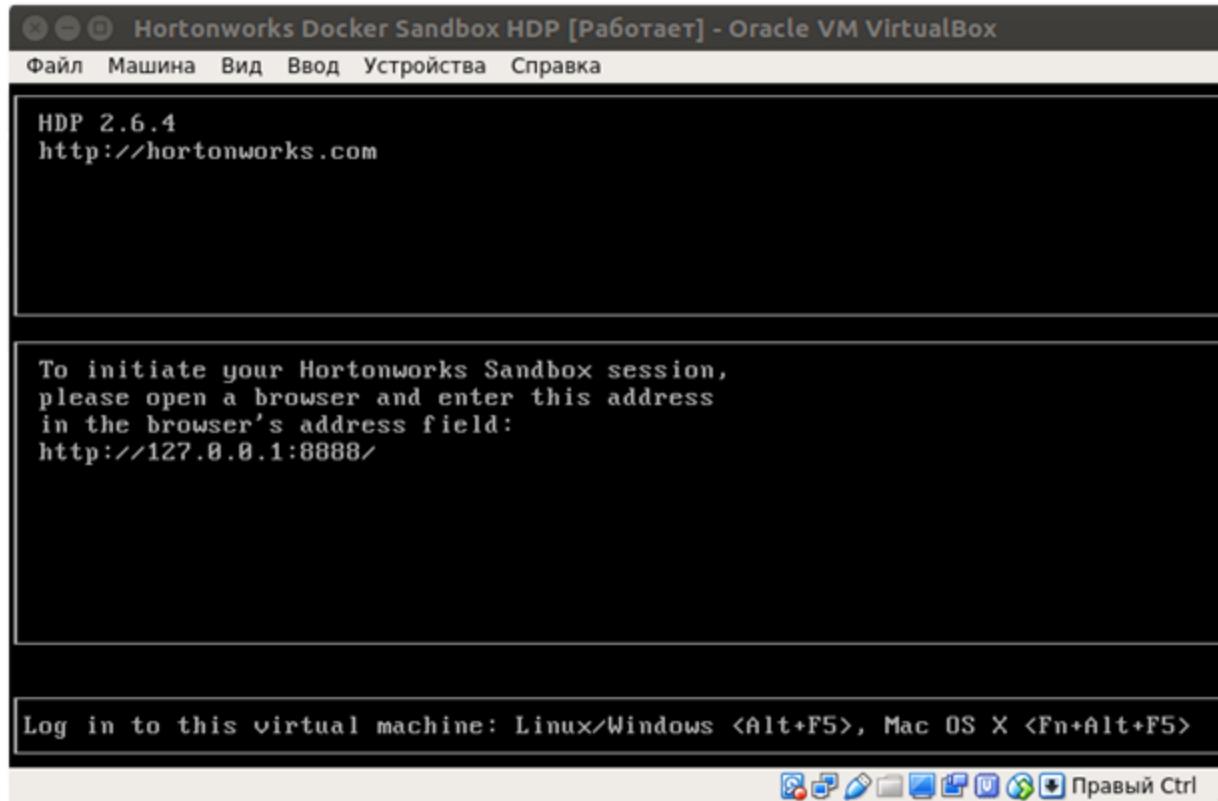


BIG DATA LANDSCAPE



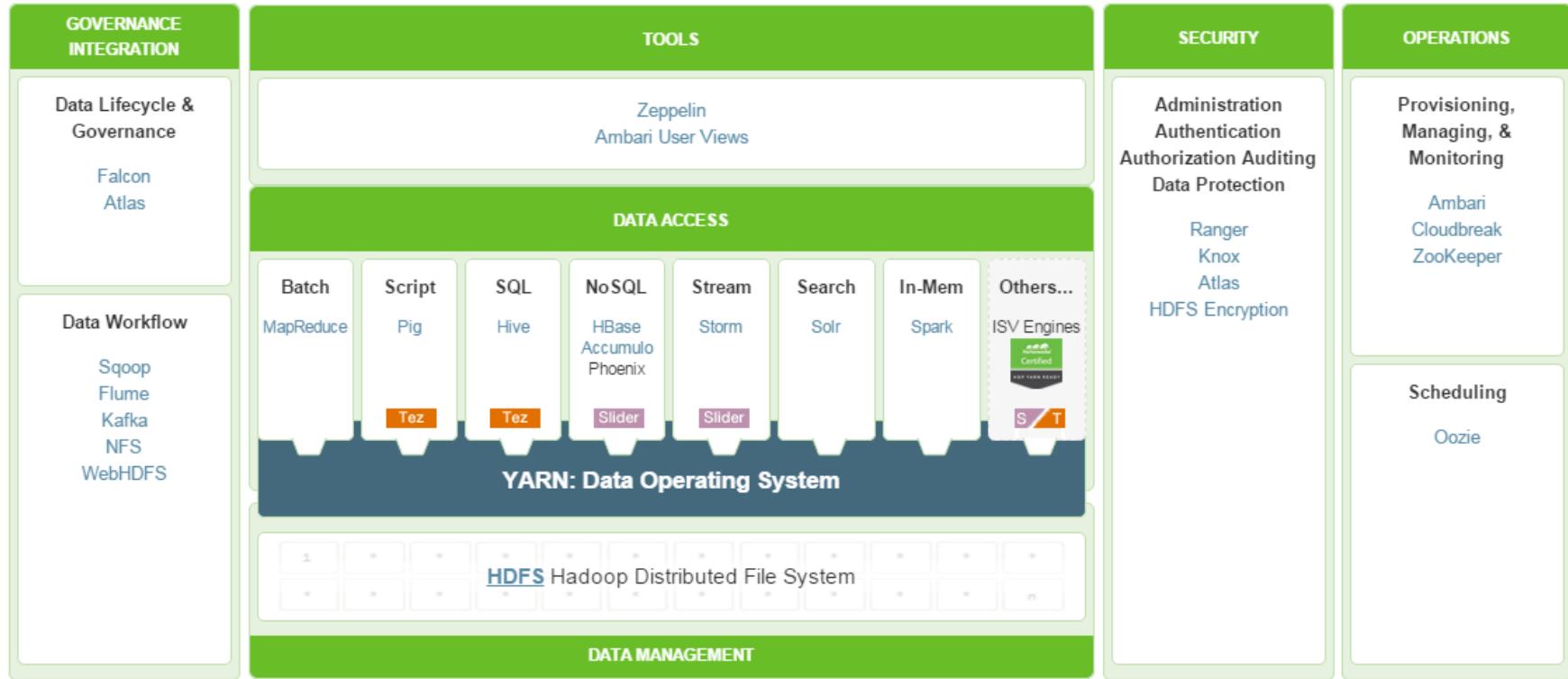


CHECK YOURSELF



HORTONWORKS

HORTONWORKS



LET'S PLAY

The screenshot shows the Ambari Dashboard interface. On the left, there's a sidebar with a list of services: HDFS, YARN, MapReduce2, Tez, Hive, HBase, Pig, Sqoop, Oozie, ZooKeeper, Falcon, Storm, Flume, Ambari Infra, Atlas, Kafka, Knox, Ranger, Spark2, Zeppelin Notebook, and Slider. Below this is an 'Actions' dropdown. At the top, it says 'Ambari Sandbox' with '0 ops' and '0 alerts'. The main area has tabs for 'Metrics', 'Heatmaps', and 'Config History'. It includes a 'Metric Actions' dropdown and a 'Last 1 hour' time range selector. The dashboard displays several metrics in cards:

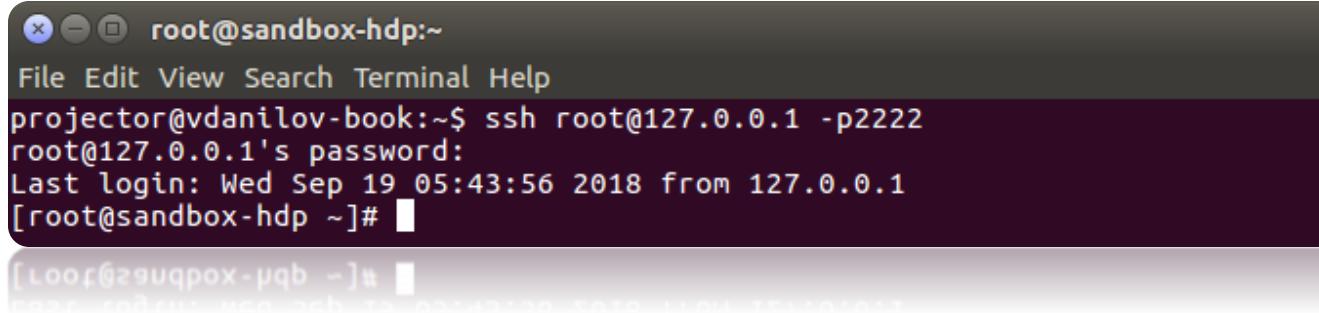
- HDFS Disk Usage: 43% (green)
- DataNodes Live: 1/1 (green)
- HDFS Links: NameNode Secondary NameNode 1 DataNodes (green)
- Memory Usage: No Data Available (grey)
- CPU Usage: No Data Available (grey)
- Cluster Load: No Data Available (grey)
- NameNode Heap: 32% (green)
- NameNode RPC: 2.00 ms (green)
- NameNode CPU WIO: n/a (grey)
- NameNode Uptime: 578.9 s (green)
- HBase Master Heap: n/a (grey)
- HBase Links: No Active Master 1 RegionServers n/a (grey)
- HBase Ave Load: n/a (grey)
- HBase Master Uptime: n/a (grey)
- ResourceManager Heap: 23% (green)
- ResourceManager Uptime: 111.9 s (green)
- YARN Memory: 0% (grey)
- NodeManagers Live: 1/1 (green)
- YARN Links: ResourceManager 1 NodeManagers (grey)

- **Login**

address: <http://127.0.0.1:8080/>
creds: admin / workshop

<http://127.0.0.1:8080>

LET'S PLAY



The screenshot shows a terminal window with the title bar "root@sandbox-hdp:~". The menu bar includes "File", "Edit", "View", "Search", "Terminal", and "Help". The command "ssh root@127.0.0.1 -p2222" was run, followed by the password "root@127.0.0.1's password:". The output shows the last login information and the prompt "[root@sandbox-hdp ~]#". Below the terminal window, there is a blurred background showing a desktop environment.

```
root@sandbox-hdp:~  
File Edit View Search Terminal Help  
projector@vdanilov-book:~$ ssh root@127.0.0.1 -p2222  
root@127.0.0.1's password:  
Last login: Wed Sep 19 05:43:56 2018 from 127.0.0.1  
[root@sandbox-hdp ~]#
```

- **SSH**

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

creds: **root / workshop**

- **Run**

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

LET'S PLAY

```
root@sandbox-hdp:~$ Sh +  
sandbox login: root  
root@sandbox.hortonworks.com's password:  
Last login: Tue Sep 18 09:57:09 2018 from 127.0  
[root@sandbox-hdp ~]#
```

- Login
address: <http://127.0.0.1:4200/>, creds: **root / workshop**

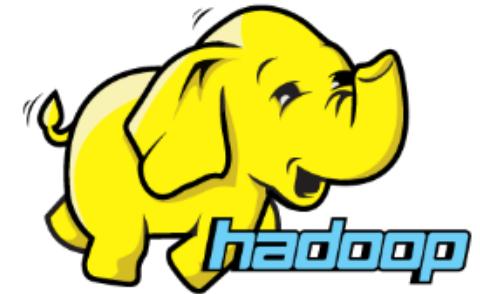
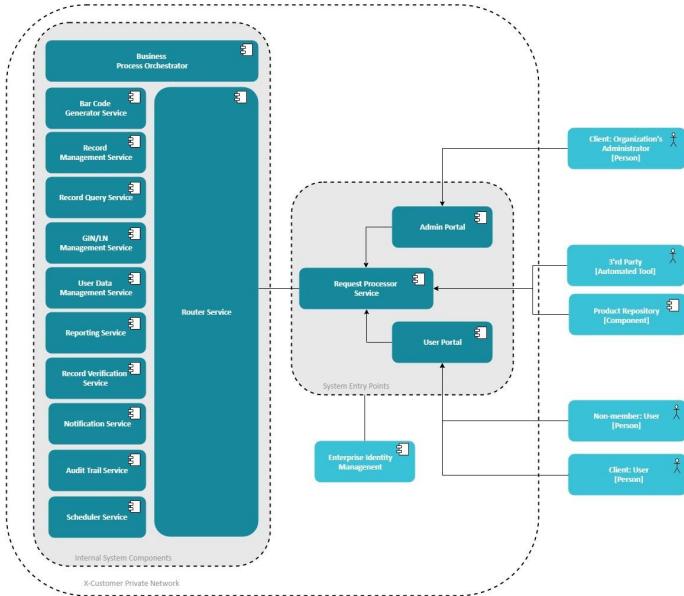
- Run

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

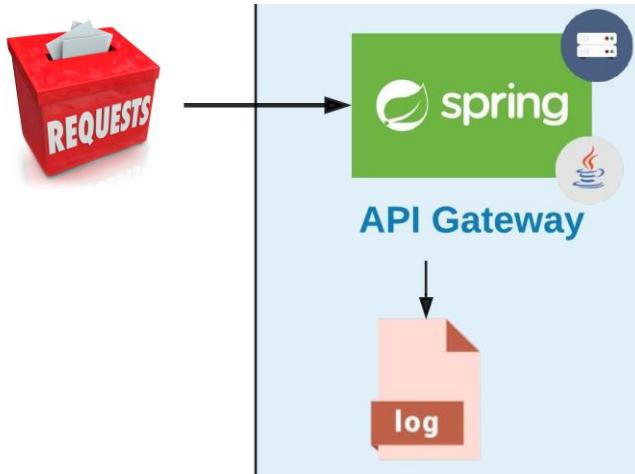
THE APPLICATION

[HTTPS://GITHUB.COM/EPAM-ZHUJ/BIGDATA-WORKSHOP](https://github.com/EPAM-ZHUJ/BIGDATA-WORKSHOP)

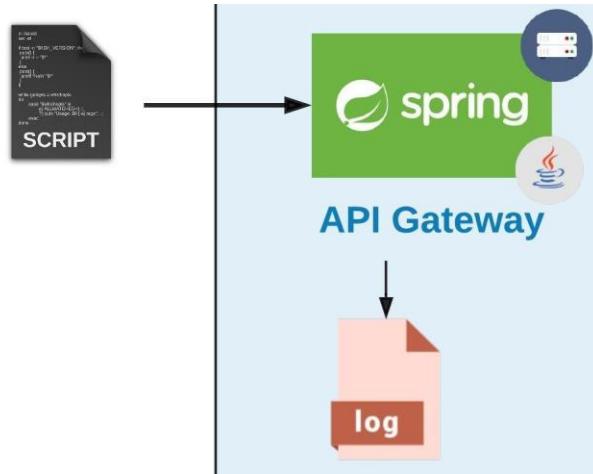
SOLUTION PRESENTATION: FOOTPRINT



PUBLIC ENDPOINT



PUBLIC ENDPOINT



LET'S INTRODUCE A COUPLE OF SERVICES



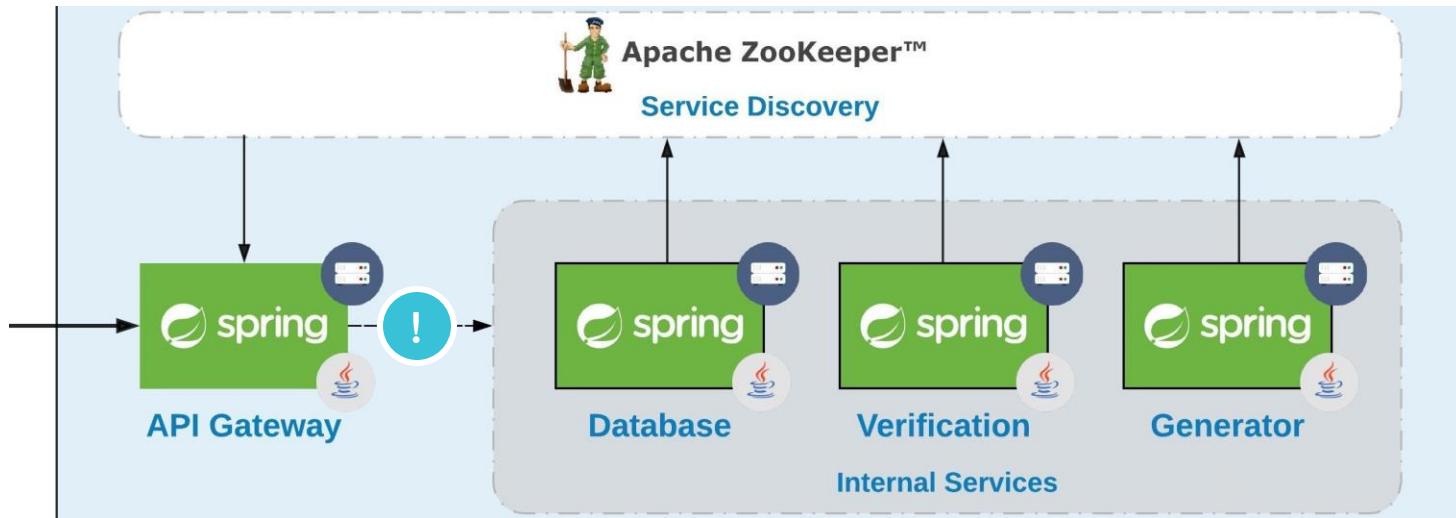
PUBLIC ENDPOINT



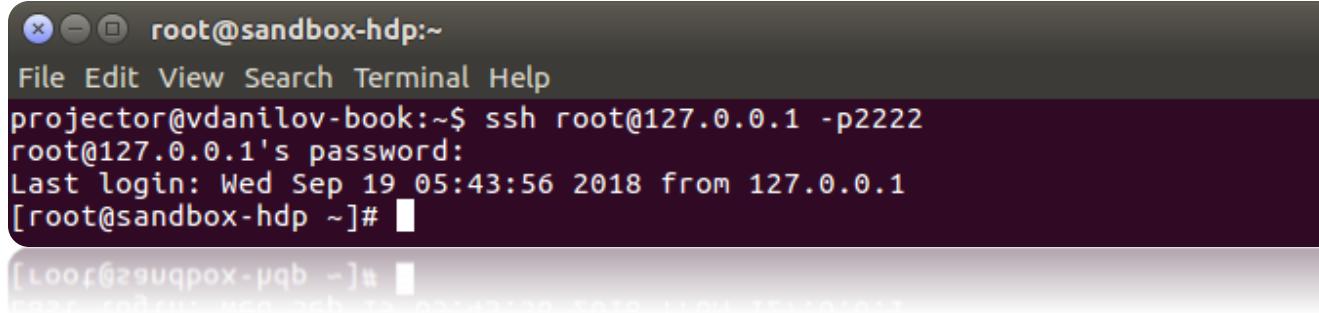
```
{  
  "userIds": ["Dmitrii", "Jennifer", "Daniel", "Olga", "Kathryn", "Sunil", "Maria", "Alexander"],  
  "numberOfRequests": 1000,  
  "workflows": [  
    {  
      "steps": [  
        {"serviceName": "database", "invocationDelay": 0},  
        {"serviceName": "verification", "invocationDelay":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



A screenshot of a terminal window titled "root@sandbox-hdp:~". The window shows a successful SSH session to a local host (127.0.0.1) on port 2222. The password was entered, and the user logged in as root. The terminal prompt is "[root@sandbox-hdp ~]#". Below the terminal window, there is some faint text from another application window.

```
root@sandbox-hdp:~  
File Edit View Search Terminal Help  
projector@vdanilov-book:~$ ssh root@127.0.0.1 -p2222  
root@127.0.0.1's password:  
Last login: Wed Sep 19 05:43:56 2018 from 127.0.0.1  
[root@sandbox-hdp ~]#
```

- **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

swagger

gateway-controller : Gateway Controller

POST /bdcc-workshop/gateway

Parameters

Parameter	Value
workflowContainer	{ "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}] }], }

Parameter content type: application/json ▾

Response Messages

HTTP Status Code	Reason	Response Model
200	OK	
201	Created	
401	Unauthorized	
403	Forbidden	
404	Not Found	

Try it out! Hide Response

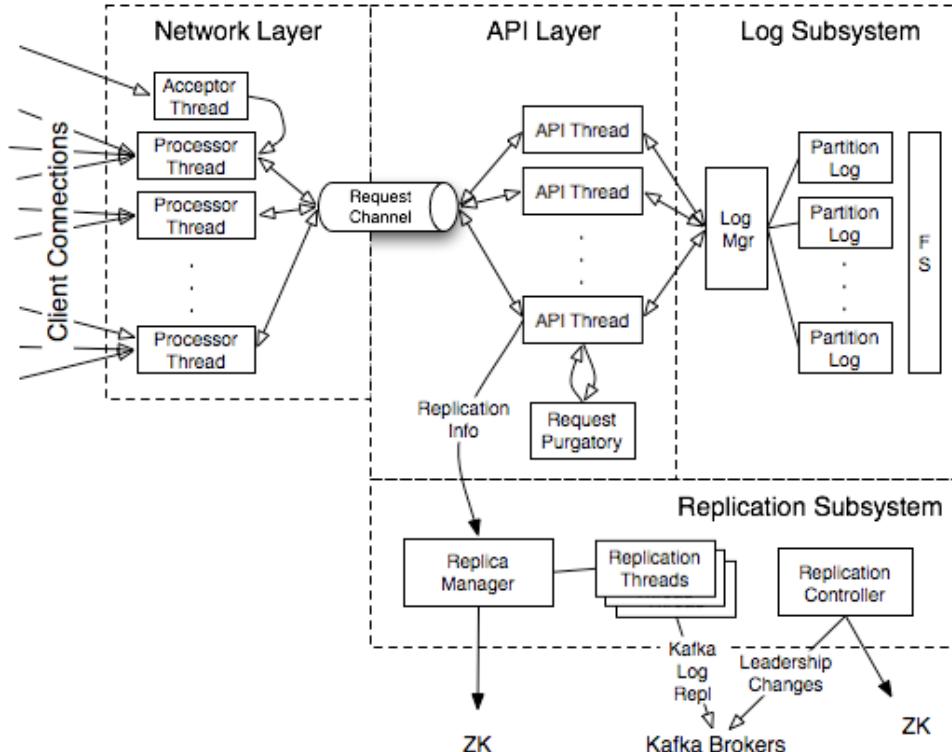
- Open <http://127.0.0.1:15505/swagger-ui.html>
- Find endpoint / method
endpoint= gateway-controller / POST
- Set script text
solution-services/.../request_template.json
- Try it out!

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

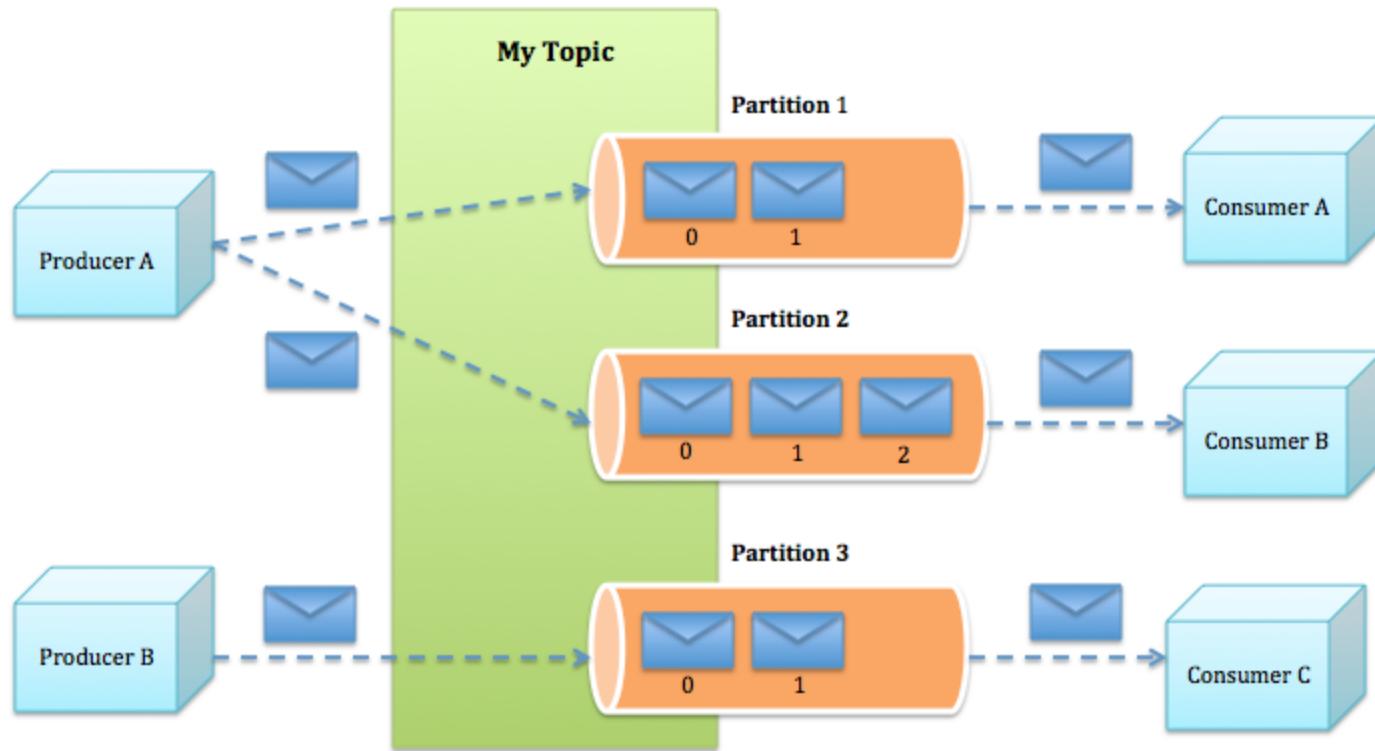
APACHE KAFKA

APACHE KAFKA

Kafka Broker Internals



APACHE KAFKA



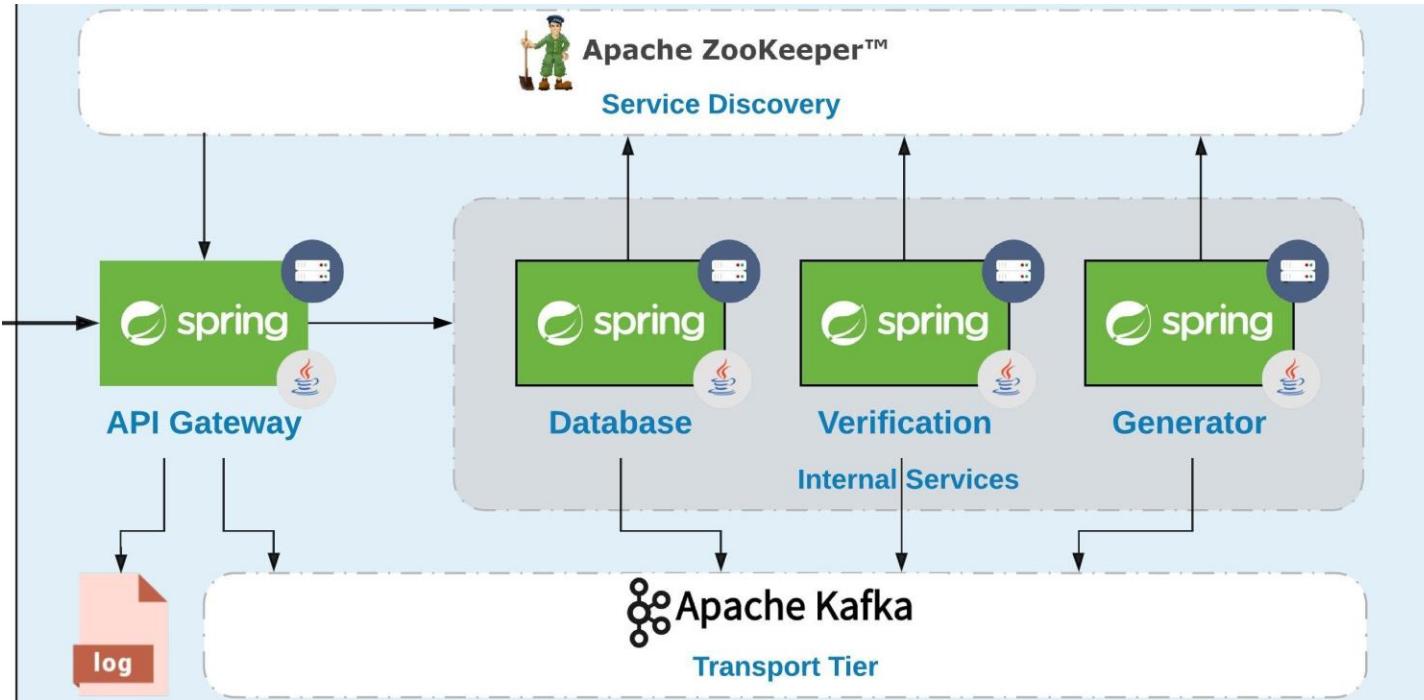
LET'S PLAY

A screenshot of a terminal window titled "root@sandbox-hdp:~\$". The window shows a root shell session on a local host. The user has run the command "ssh root@127.0.0.1 -p2222" and is prompted for the password. The password entered is "root". The terminal then displays the last login information and the prompt "[root@sandbox-hdp ~]#".

```
root@sandbox-hdp:~$ +  
san x - san root@sandbox-hdp:  
File Edit View Search Terminal Help  
Lasprojector@vdanilov-book:~$ ssh root@127.0.0.1 -p2222  
[root@127.0.0.1's password:  
Last login: Wed Sep 19 05:43:56 2018 from 127.0.0.1  
[root@sandbox-hdp ~]#
```

- 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

```
root@sandbox-hdp:~$ Sh +  
sandbox login: root  
root@sandbox.hortonworks.com's password:  
Last login: Tue Sep 18 09:57:09 2018 from 127.0  
[root@sandbox-hdp ~]#
```

Fri Sep 21 2018 10:00:00 AM UTC

- **Login**
address: <http://127.0.0.1:4200/>, creds: **root / workshop**
- **Run**

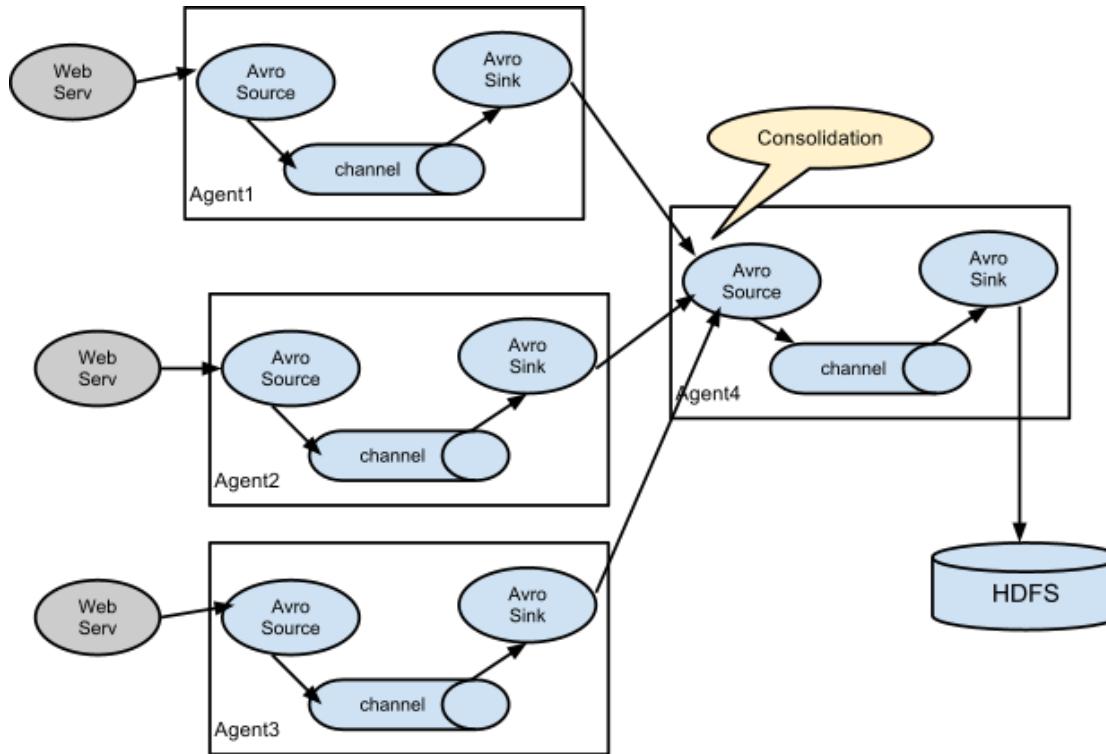
```
# cd workshop/kafka
# ./kafka-console-consumer.sh --topic resource.utilization.topic
--bootstrap-server 0.0.0.0:6667
```
- **Have fun (with the App)**

LET'S PLAY

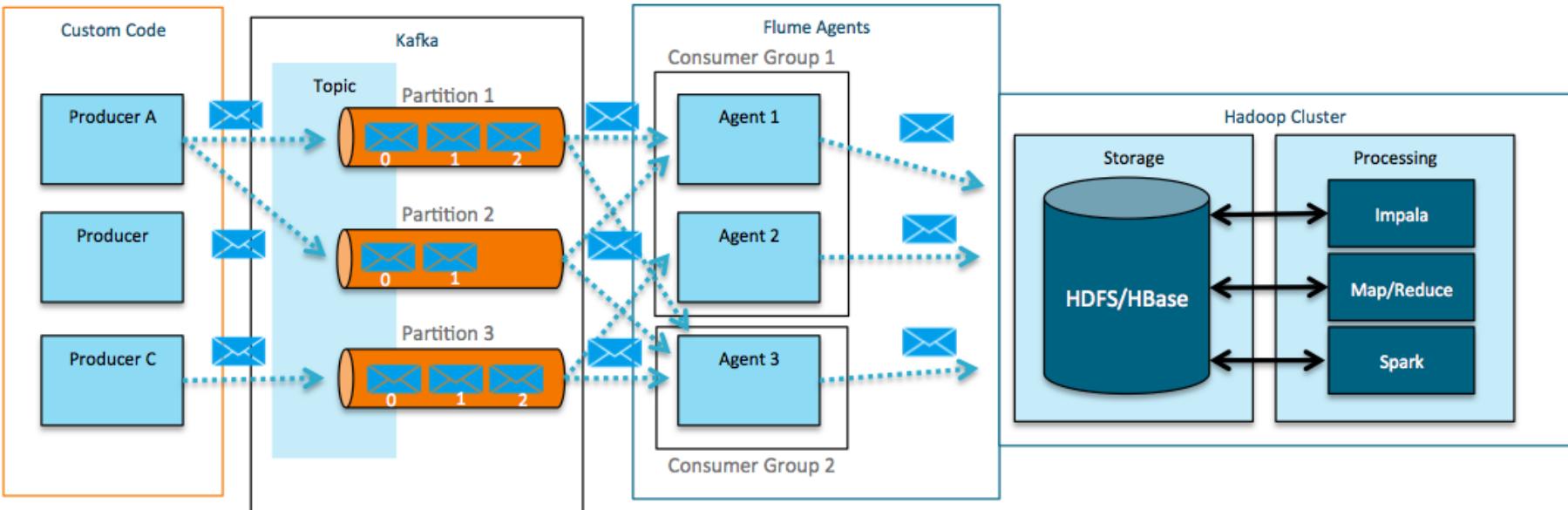
```
generator] (2018-09-10 16:53:00.569, 2018-09-10 16:53:19.765) | Daniel | 0097600c-0107-4cc7-0507-104407100e7d | avg cpu time: 0.146699286600234
generator] (2018-09-18 18:53:08.569, 2018-09-18 18:53:11.403) | Alexander | aa9d362e-92e5-4034-b05a-ca8991932eba | avg cpu time: 0.647632191968
generator] (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-8666041a9715b | 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: 0.14426436345754
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: 0.0325239650447052
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 | 24cefdfc-cble-4b75-9d6d-5f3f8281d007 | putSize: 933; returnSize: 1585
database] 2018-09-18 18:53:09.824 | Daniel | 001c59a2-d38c-4f28-98ae-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 | 24cefdfc-cble-4b75-9d6d-5f3f8281d007 | records verified: 38769
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: 0.6650131410260020
generator] (2018-09-18 18:53:11.830, 2018-09-18 18:53:16.987) | Olga | 24cefdfc-cble-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-4f28-98ae-8e440bc64101 | avg cpu time: 0.5094041136983717
database] 2018-09-18 18:53:12.827 | Kathryn | bd8bdff8c-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 | ccff86d2c-83b9-40ef-9a48-94a02a1da808 | putSize: 673; returnSize: 2766
verification] 2018-09-18 18:53:13.845 | Kathryn | bd8bdff8c-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 | ccff86d2c-83b9-40ef-9a48-94a02a1da808 | records verified: 1485
generator] (2018-09-18 18:53:14.862, 2018-09-18 18:53:16.052) | Kathryn | bd8bdff8c-8836-494d-bbb7-c25920ebd9bc | avg cpu time: 0.223237676403219
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
generator] (2018-09-18 18:53:14.900, 2018-09-18 18:53:16.010) | Sunil | ccff86d2c-83b9-40ef-9a48-94a02a1da808 | avg cpu time: 0.5623615384053047
```

APACHE FLUME

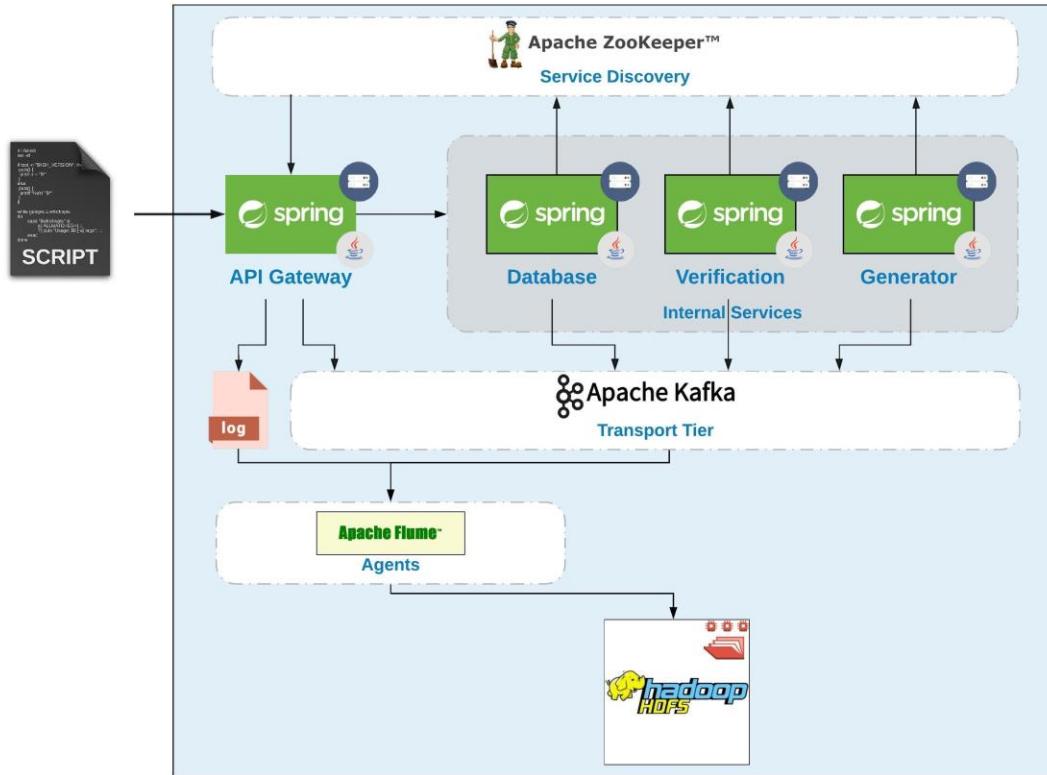
APACHE FLUME



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

```
root@sandbox-hdp:~ Sh +  
sandbox login: root  
root@sandbox.hortonworks.com's password:  
Last login: Tue Sep 18 09:57:09 2018 from 127.0  
[root@sandbox-hdp ~]#
```

1. Logon to your machine

- **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

The screenshot shows the Ambari UI interface. At the top, there is a navigation bar with the Ambari logo, the word "Sandbox", and status indicators for "0 ops" and "0 alerts". On the right side of the top bar is a user dropdown menu labeled "admin". Below the top bar, there is a breadcrumb navigation path: "/ > user > workshop > resource.utilization.topic...". To the right of the path, it says "Total: 2 files or". On the left, there are three small icons: a house, a document, and a circular arrow. The main content area displays a table with two rows of data. The columns are labeled "Name >", "Size >", and "Last Modified >". The first row contains the file "FlumeData.1537200181504" with a size of "226.4 kB" and a last modified date of "2018-09-17 19:13". The second row contains the file "FlumeData.1537297433743" with a size of "16.2 kB" and a last modified date of "2018-09-18 22:06". Below the table, there is some very faint, illegible text.

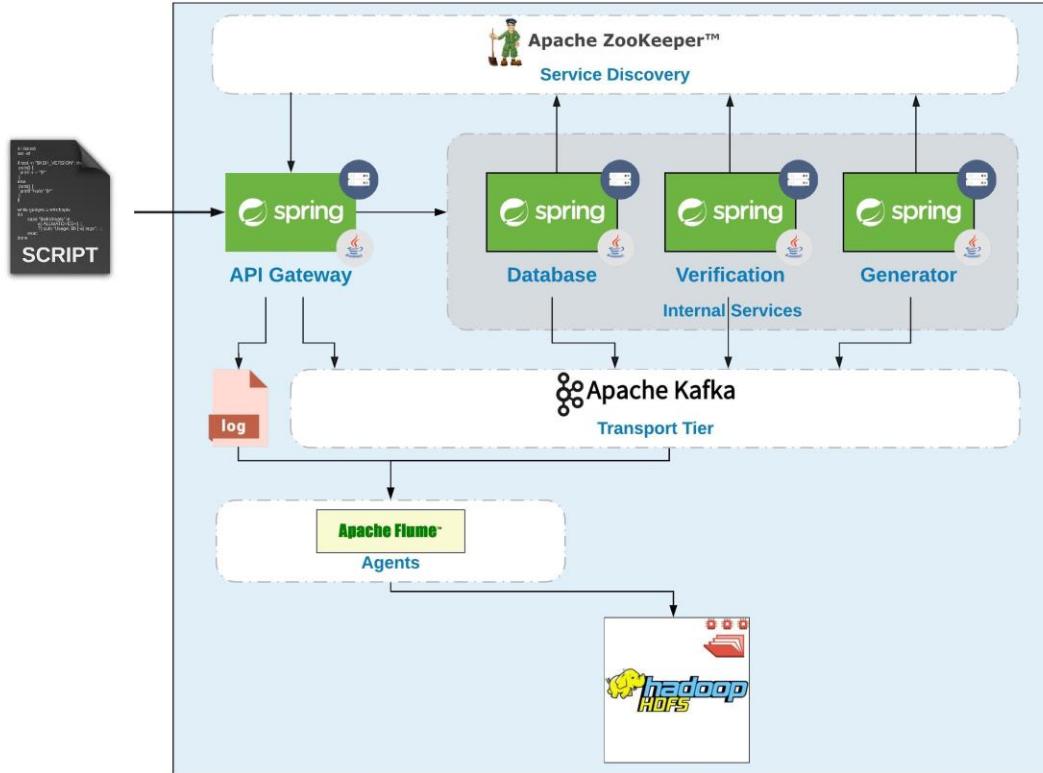
Name >	Size >	Last Modified >
FlumeData.1537200181504	226.4 kB	2018-09-17 19:13
FlumeData.1537297433743	16.2 kB	2018-09-18 22:06

<http://127.0.0.1:8080>

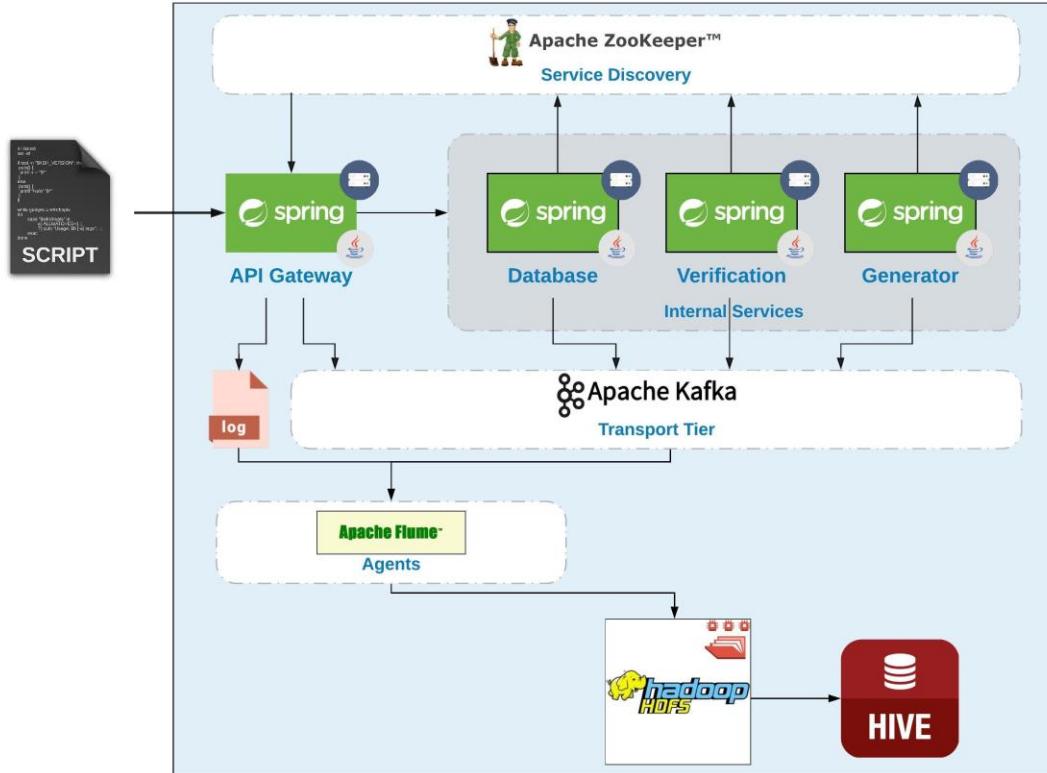
http://127.0.0.1:8080/#/main/view/FILES/auto_files_instance

MAP-REDUCE

APPLICATION

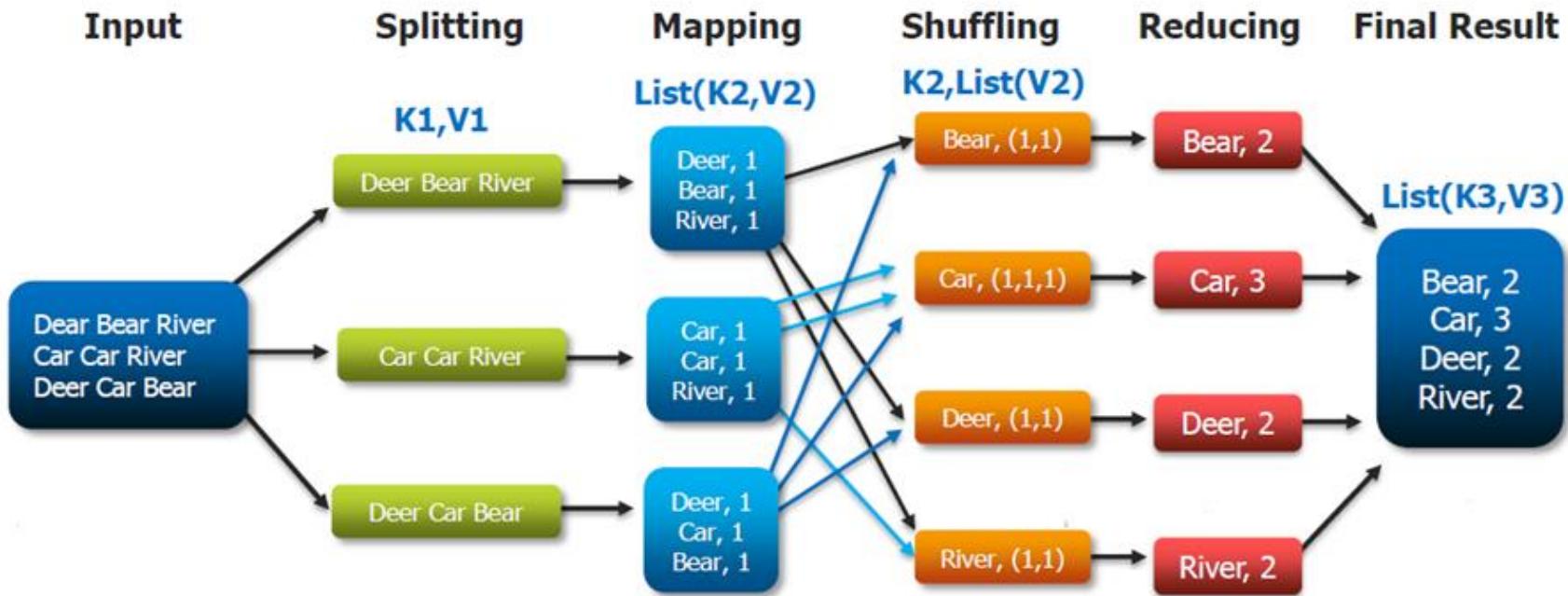


APPLICATION



MAP-REDUCE

The Overall MapReduce Word Count Process



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
```

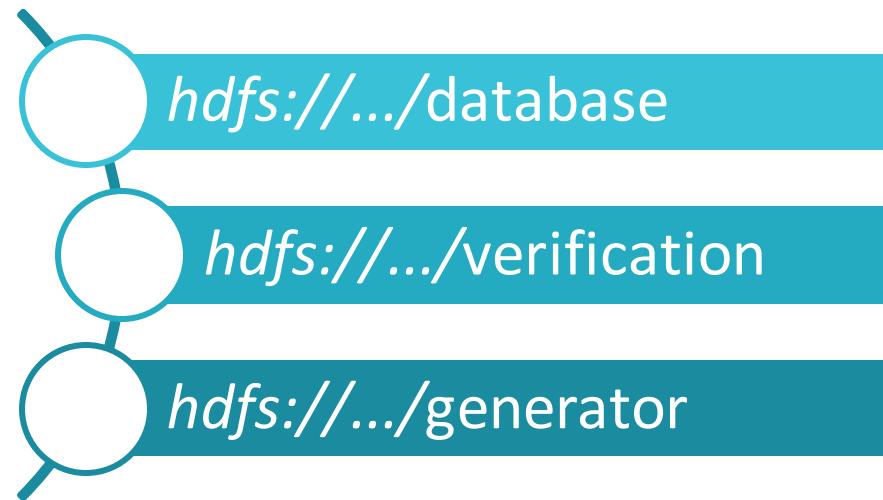
generator-service

```
"[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

```
root@sandbox-hdp:~$ Sh +  
sandbox login: root  
root@sandbox.hortonworks.com's password:  
Last login: Tue Sep 18 09:57:09 2018 from 127.0  
[root@sandbox-hdp ~]#
```

1. Logon to your machine

- **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/172.17.0.2:16200
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 application for job: job_1537328875388_0002
18/09/19 04:10:00 INFO mapreduce.YarnClientImpl: Submitting 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 04:10:07 INFO mapreduce.Job: Running job: job_1537328875388_0002
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 04:10:14 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: 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=365967
  Map output materialized bytes=365967
```

```
  Input split bytes=338
  Combine input records=0
  Combine output records=0
  Reduce input groups=11
  Reduce input records=365967
  Reduce output records=2181
  Reduce output records=0
  Spilled Records=4362
  Shuffled Maps=0
  Failed Shuffles=0
  Merged Map outputs=2
  GC time elapsed (ms)=736
  CPU time spent (ms)=2660
  Physical memory (bytes) snapshot=5452086272
  Virtual memory (bytes) snapshot=5405017680
  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=365961
  Bytes Written=0
  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

The screenshot shows the Ambari UI interface for managing HDFS files. At the top, there's a navigation bar with the Ambari logo, a 'Sandbox' link, and status indicators for '0 ops' and '0 alerts'. On the right, a user dropdown menu is open, showing 'admin' and several links: YARN Queue Manager, Files View (which is highlighted in yellow), Hive View, Hive View 2.0, Pig View, Storm View, Tez View, and Workflow Manager.

The main area displays a file browser for the 'workshop' directory. The path is shown as '/ > user > workshop'. Below the path are standard file operations: Open, Rename, Permissions, Delete, Copy, Move, Download, and concatenate. A message '0 Files, 1' is displayed in a yellow box.

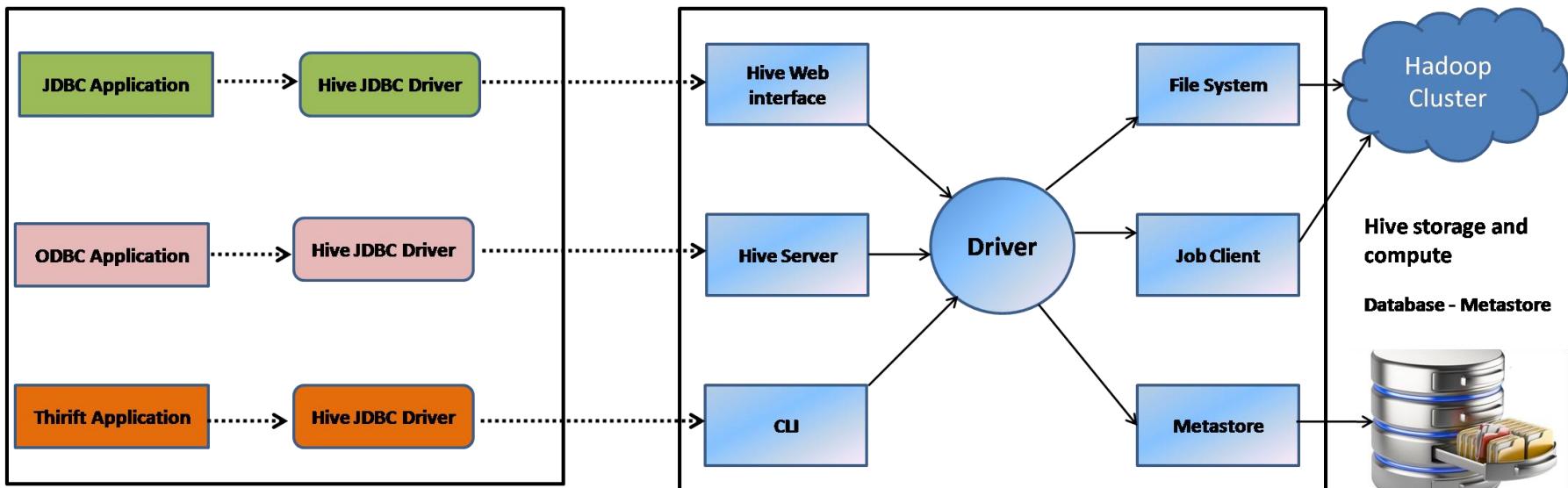
Name	Size	Last Modified
hive_data	--	2018-09-19 06:53
resource.utilization.topic	--	2018-09-18 22:06

At the bottom left, there's a link to 'resource.utilization.php'. On the right side of the page, two URLs are listed:

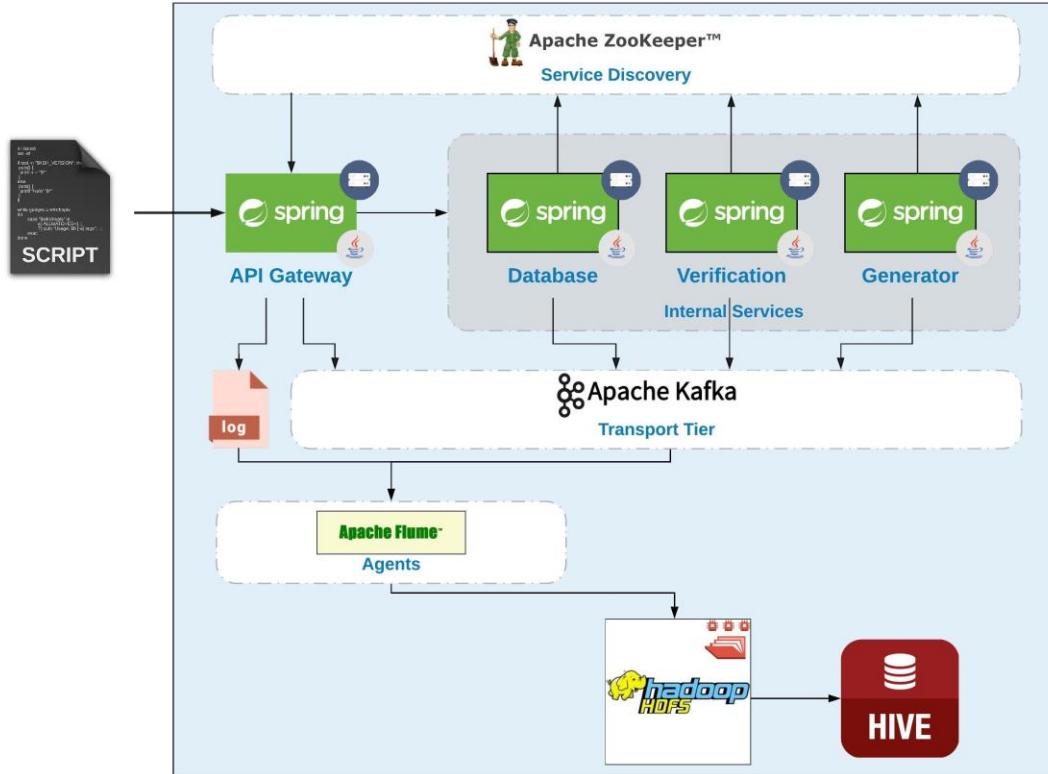
- <http://127.0.0.1:8080>
- http://127.0.0.1:8080/#/main/view/FILES/auto_files_instance

HIVE

Hive Architecture



APPLICATION



LET'S PLAY

The screenshot shows the Ambari Hive interface. At the top, there's a navigation bar with icons for Ambari, Sandbox, 0 ops, and 0 alerts. Below the navigation bar, there are tabs for Hive, Query, Saved Queries, History, UDFs, and Upload Table. A dropdown menu for the user 'admin' is open, showing options like YARN Queue Manager, Files View, Hive View (which is selected), Hive View 2.0, Pig View, Storm View, Tez View, and Workflow Manager.

The main area has two sections: 'Database Explorer' on the left and 'Query Editor' on the right. In the Database Explorer, the 'default' database is selected. The 'Databases' section lists several databases: default, database service stats, generator service stats, sample 07, sample 08, users, verification service stats, foodmart, and xademo. Below the databases, there are sections for 'Tables', 'Partitions', 'Views', and 'UDFs'. In the Query Editor, the 'Worksheet' tab is active. It contains the following SQL code:

```
1 DROP TABLE IF EXISTS verification_service_stats
2 CREATE EXTERNAL TABLE verification_service_stat
3   event timestamp TIMESTAMP,
4   user_id STRING,
5   workflow_id STRING,
6   records_verified BIGINT)
7 ROW FORMAT DELIMITED
8 FIELDS TERMINATED BY '\001'
9 LOCATION '/user/workshop/hive_data/verification';
```

Below the code, there are four buttons: Execute (green), Explain, Upload, and Save as... .

At the bottom, there's a section titled 'Query Process Results (Status: SUCCEEDED)' which displays the status of the query execution.

<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

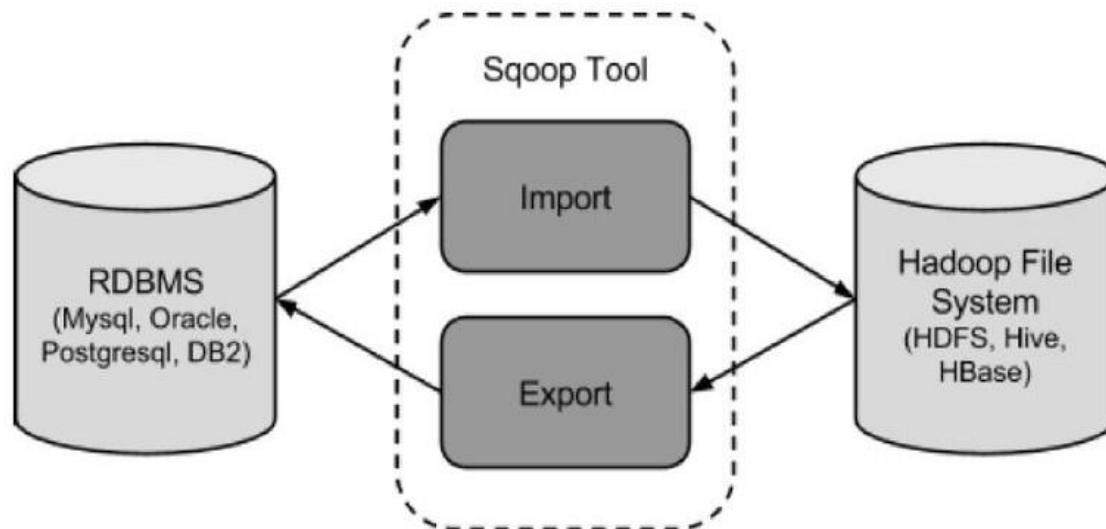
with

```
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;
```

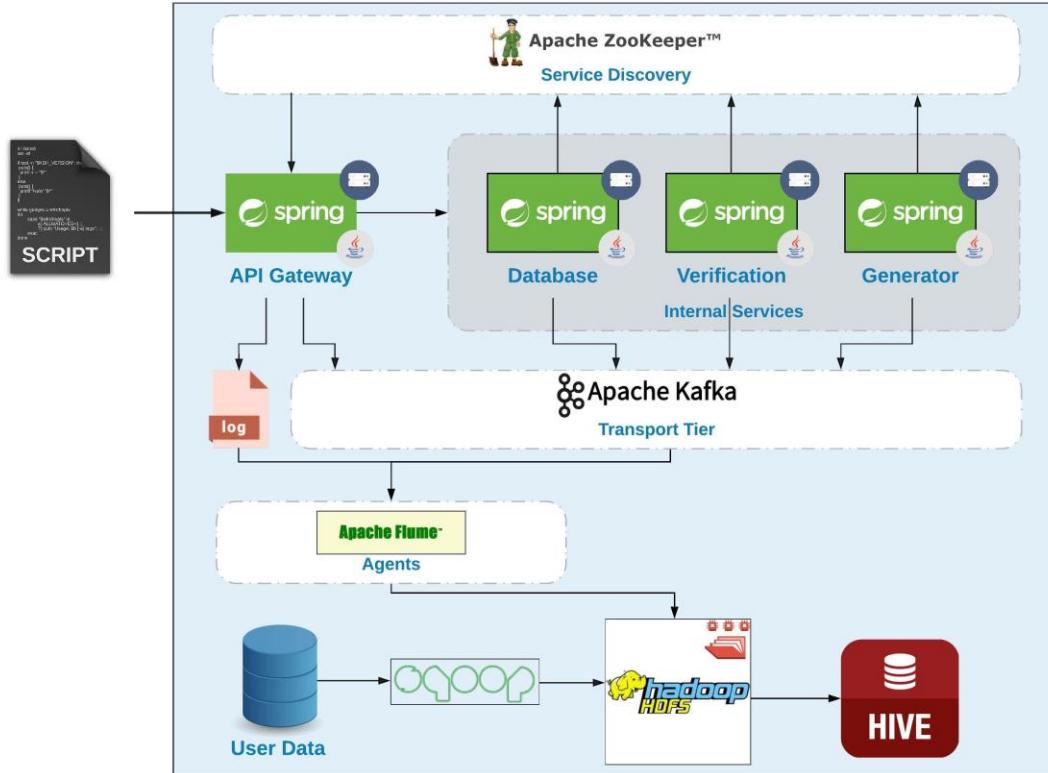
hive/top_database_disk_consumption_user.txt

SQOOP

SQOOP



APPLICATION



LET'S PLAY

```
root@sandbox-hdp:~$ Sh +  
sandbox login: root  
root@sandbox.hortonworks.com's password:  
Last login: Tue Sep 18 09:57:09 2018 from 127.0  
[root@sandbox-hdp ~]#
```

- **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-separated-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 should 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 04:46:45 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM users AS t WHERE 1=0
18/09/19 04:46:45 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM users AS t WHERE 1=0
18/09/19 04:46:45 INFO org.CompilationManager: HADOOP_MAPRED_HOME is /usr/hdp/2.6.4.0-91/hadoop-mapreduce
Note: /tmp/sqoop-root/compile/4e3268febe518ea94463f8248f3bfb3/users.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
18/09/19 04:46:47 INFO org.CompilationManager: Writing jar file: /tmp/sqoop-root/compile/4e3268febe518ea94463f8248f3bfb3/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 04:46:48 INFO client.RMProxy: Connecting to ResourceManager at sandbox-hdp.hortonworks.com/172.17.0.2:8032
18/09/19 04:46:48 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 committed 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 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=8
    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

Filler 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

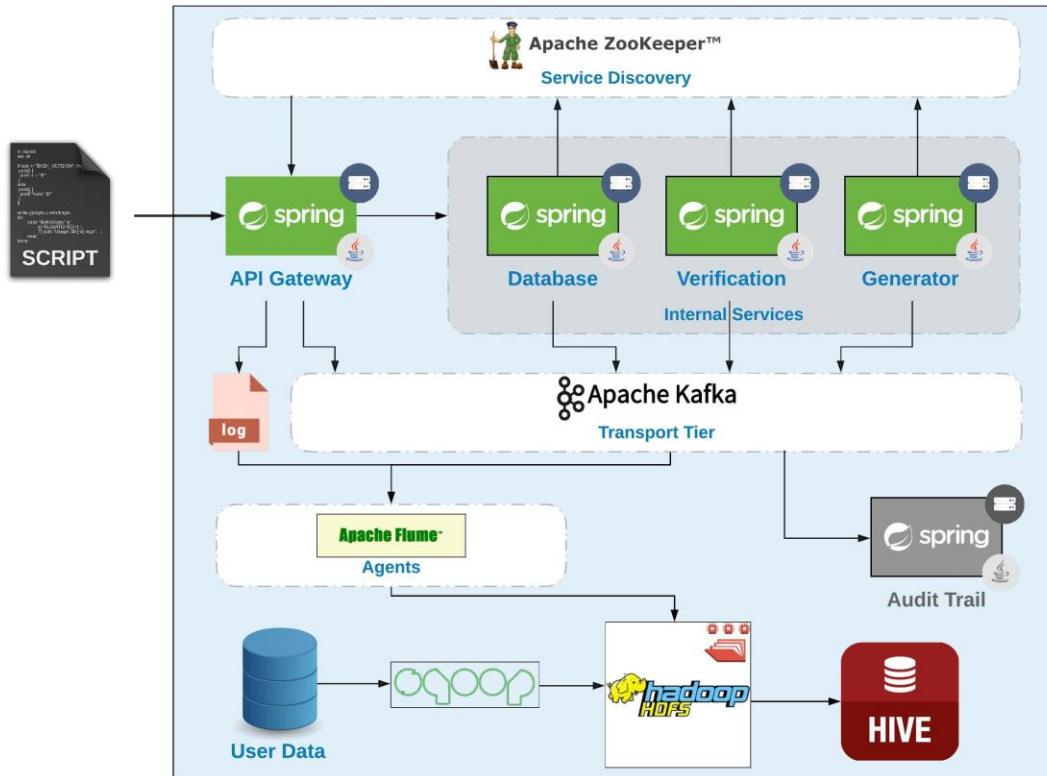
LET'S PLAY

```
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 swo.org_summed_sizes desc) as ranked_org_summed_sizes,
swo.* from summed_by_organizations swo
```

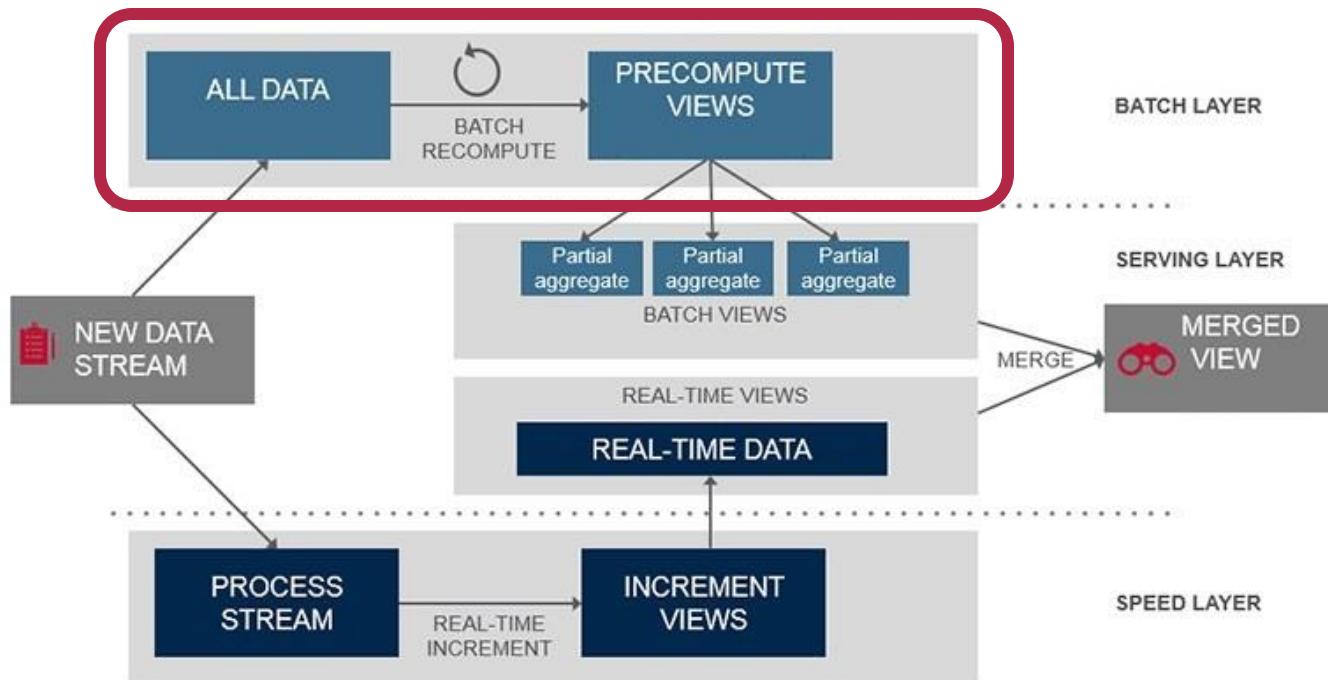
hive/top_database_disk_consumption_organization.txt

YOUR TURN

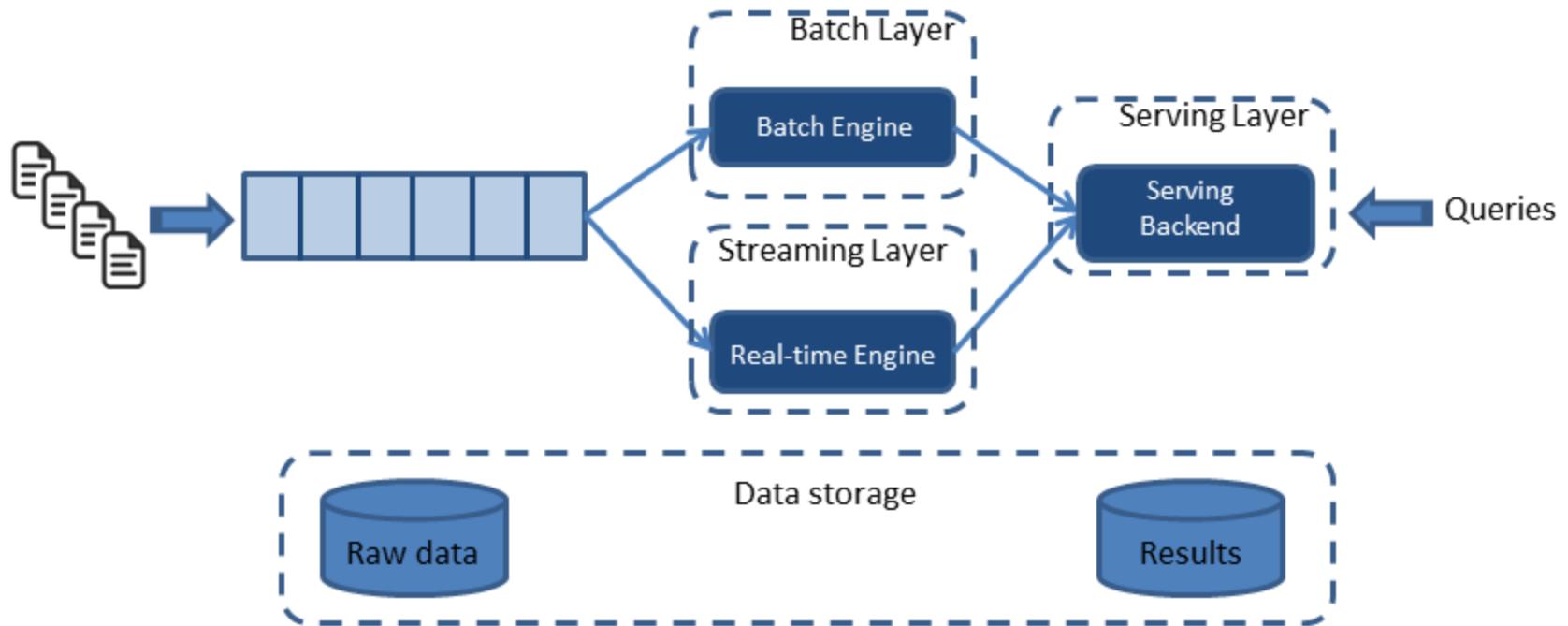
REALTIME LAYER? YOUR TURN ;)



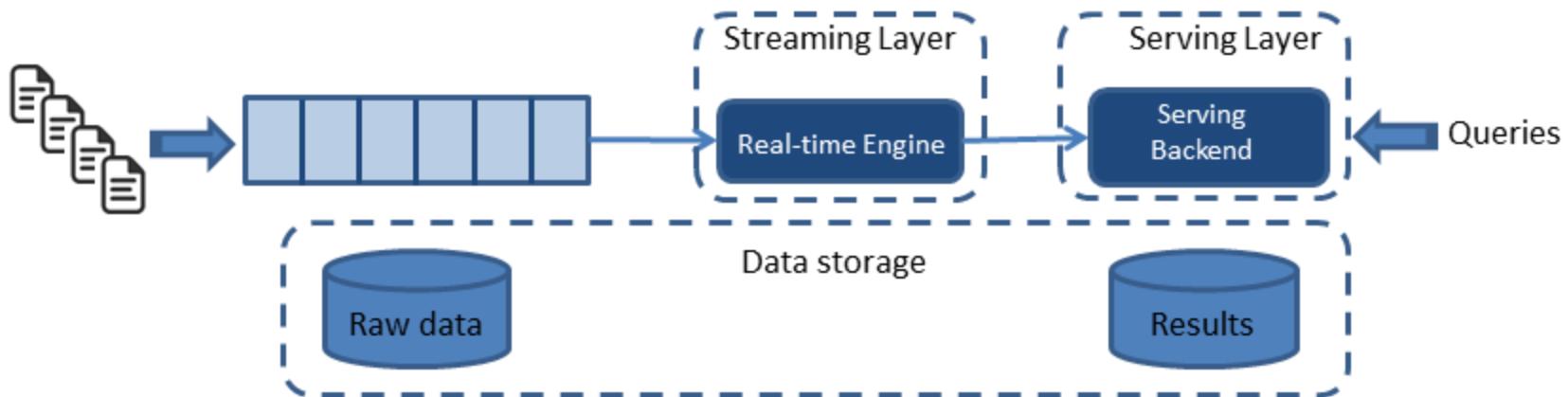
BIG DATA APPROACH: LAMBDA ARCHITECTURE



BIG DATA APPROACH: LAMBDA ARCHITECTURE



BIG DATA APPROACH: KAPPA ARCHITECTURE



A group of business people in a modern office lobby. In the foreground, a man in a dark suit and tie is shaking hands with another man in a light blue suit. They are standing at a glass table with documents and pens on it. Other people are visible in the background, some looking at a clipboard. The ceiling has a large, modern architectural structure.

THANK YOU!