Emergence of Real-Time: Real-time Analysis of Customer Financial Activities With Apache Flink

Emerging Technology For Enterprise Conference
April 11 - 12, 2016
Philadelphia, PA

Srinivas (Srini) Palthepu, Ph.D.
Sr. Manager, Big Data Engineering
Capital One
1. Capital One
2. Traditional Batch Analytics
3. The Great Paradigm Shift – Real-Time Analytics
4. What are the Drivers?
5. Apache Flink – Next Generation Big Data Analytics Framework
6. Business Use Case: Customer Activity Event Logs
7. Conclusions
Capital One is a *software engineering* company *whose products happen to be financial products*

- First Bank to go to Cloud
- First Bank to Contribute to Open Source
- First Bank to Support Technology Community Engagement
- Driving the innovation and technology, not just consumers

**Embracing Open Source with strategic purpose, not just the cost!**
Agenda

1. Capital One
2. Traditional Batch Analytics
3. The Great Paradigm Shift – Real-Time Analytics
4. What are the Drivers?
5. Apache Flink – Next Generation Big Data Analytics Framework
6. Business Use Case: Customer Activity Event Logs
7. Conclusions
2. Traditional Batch Analytics

1. Traditional Batch Analytics Architecture
2. What is CSAD Cycle?
3. Limitations of Traditional Approach
2.1 Traditional Batch Analytics

- Operation Store
- ETL
- Warehouse
- Sandbox
- Datamarts

Actions based on Insights
2.2 What is CSAD Cycle?

- Application generates data that is Captured into operational store
- Periodically move the data (typically daily) to some data processing platform and run ETL to clean, transform, enrich data
- Load the data into various places for various uses such as Warehouse, OLAP cubes, Marts
- Use Analytics Tools such as R, SAS, SQL, or Dashboard/Reporting tools to find insights
- Decide what actions can be implemented based on the insights
2.3 Limitations of Traditional Batch Analytics

- **Time-To-Insight** is long, several days
- Spend **several days** just to get the right data in right place
- Not suited for today's business practices
- This model has not changed even after Big Data revolution!
1. Capital One
2. Traditional Batch Analytics
3. The Great Paradigm Shift – Real-Time Analytics
4. What are the Drivers?
5. Apache Flink – Next Generation Big Data Analytics Framework
6. Business Use Case: Customer Activity Event Logs
7. Conclusions
3. The Great Paradigm Shift – Real-Time Analytics

1. What is Fast Data and how is it different from Big Data?
2. What is Real-Time v/s Batch – explained
3. What is Real-Time Analytics?
4. Some Real-Time Use Cases
3.1 What is Fast Data?

- **Fast Data** is a new buzzword that is slowly overtaking Big Data.
- Big Data is characterized by 3 V (Volume, Variety, and Velocity). Much of the last decade with Hadoop is focused on storing and processing large volumes of data in batch-oriented fashion.
- **Fast Data** is characterized by processing of large amounts of data coming at **High Speed** that needs to be processed continuously and acted upon in real-time.
- Real-Time data processing is characterized by **Unbounded Data**.
- **High-Speed** and **Low-Latency** is name of the game!
- Depending upon use case, sometimes Latency is less important than semantics and capabilities.
3.2 Real-Time v/s Batch – Water Heater

- **Batch Water Heater**
  - Collect water into the tank
  - Heat the water in the tank (process)
  - Supply water **after the water is heated**
  - Wait till the **whole batch** to heat to desired level
  - Heating may be continuous, but the supply is batch

**Store - Process - Serve** Model
3.2 Real-Time v/s Batch – Water Heater

Real-Time Water Heater
- Heats the water on-the-fly
- No Need to wait for hot water (low-latency)
- Capacity of heater to match the volume and velocity of flow
3.3 Real-Time Analytics

- **Real-Time Analytics** aims to reduce the traditional CSAD cycles to minimum, **few seconds**, sometimes **sub-second**.

- **Problems with traditional Batch Analytics:**
  - Old data, often **stale**
  - Too slow for fast paced world
  - Need to act sooner, sometimes instantly based on customer behavior

- **Real-Time Analytics** will address these issue associated with Batch Oriented Traditional Analytics
3.4 Real-Time Analytics – Use Cases

Use Cases From Financial World

➢ Real-Time Fraud Prevention
  - Detect fraudulent transaction on the fly rather than after the transaction is approved

➢ Second-Look of duplicate transaction
  - Point of Sales Error, Duplicate Charges detected before you leave the store!

➢ Real-Time CLIP Decision
  - Credit Limit Increase on-the-fly when a transaction pushes above the limits

➢ Real-Time Targeted offers
  - Special offers pushed to user based on users real-time information location, status and earlier actions.

➢ Real-Time Customer Assistant
  - Detect what customer is trying to do and intervene in real-time

➢ Real-Time Shopping Advice
3.4 Real-Time Analytics – Use Cases

Other Use Cases

- **Internet of Things (IoT)**
  - Streaming sensor data analyzed real-time and acted-upon

- **Real-Time System Monitoring and Failure Prevention**
  - Failure Never Happen Suddenly – There are early warnings!

- **Connected Automobiles**
  - Airbus has 10000 sensors
  - Constant Monitoring and feedback. Continuous Learning of driver’s behavior

- **Health Monitoring Medical Devices**
Agenda

1. Capital One
2. Traditional Batch Analytics
3. The Great Paradigm Shift – Real-Time Analytics
4. What are the Drivers?
5. Apache Flink – Next Generation Big Data Analytics Framework
6. Business Use Case: Customer Activity Event Logs
7. Conclusions
4. What are the Drivers?

1. Business Drivers
   - Business Environment became very competitive
   - Need to act quickly for fast changing market place & consumer behavior

2. Technology Drivers
   - New Technologies enabling possibilities that were not present earlier

3. Social Behaviors
   - Consumers wants and expectations are changing fast
   - Businesses need to react to their expectations.

4. New Industries and New Use Cases
   - IoT - Internet of Things
   - Connected Automobiles
4.1 Business Drivers

- Business Environment has became very competitive
- Need to act quickly for fast changing market place & consumer behavior
- Customer Expectations
4.2 Technology Driver

- Legacy Big Data (Hadoop) solely focused on Batch Oriented Data Warehousing.
  - More Data (Volume)
  - Enabled More Types of Data (Variety)
  - More Speed (Velocity)
    - Did not change traditional CSAD cycle!

- Advancement in Big Data and Fast data is fueling a new paradigm shift
  - Apache Storm started the trend
  - Apache Spark paved the way
  - Apache Flink is taking Real-Time processing to whole new level
    - True Real-Time Stream processing (event-at-time) at scale
    - High-Performance
    - Distributed
    - Fault-Tolerant
4.2 Technology Drivers

- New Generation of Technologies such as Apache Flink can deliver **Analytics** and **Business Intelligence** in real-time.
- Businesses Need To React **Quickly** for real-world events. Can not wait for long CSAD Cycles.
- Data is becoming **obsolete** as fast as it is generated.
- **Fast Data** is like **Fast Food** : consume it quickly or it will be **stale**.
4.3 Social Trends

Oral-B®
PRO 5000 SmartSeries
Real-Time Feedback.
Superior Clean.*
4.4 New Industries and New Use Cases

• **Internet of Things (IOT) and Sensor Generated Data**
  – Every Device Is A Smart Device
  – Home Appliances

• **Connect Automobile**
  – Boeing Aircraft has 10000 sensors constantly sending the data
  – Passenger Cars are Data Generators in way that was seen never before!
4.3 Social Trends

- We all live in the world of instant gratification!
- Spread of Smartphones are raising expectations from users
  - I want everything!! and I want it now!!
- Even a simple query may need to process tons of data
  - Think about Google Translate on a smart phone!
- Emergence of Powerful Smart Phones and Mobile Computing
  - We want Everything! We Want it Now!!
Agenda

1. Capital One
2. Traditional Batch Analytics
3. The Great Paradigm Shift – Real-Time Analytics
4. What are the Drivers?
5. Apache Flink – Next Generation Big Data Analytics Framework
6. Business Use Case: Customer Activity Event Logs
7. Conclusions
5. Apache Flink – Next Generation Big Data Analytics Framework

1. What is Apache Flink
2. Flink – Next Generation Analytics Framework
3. Flink Stack
### 5.1 Apache Flink as the Next Generation of Big Data Analytics

<table>
<thead>
<tr>
<th>✓ Batch</th>
<th>✓ Batch</th>
<th>✓ Batch</th>
<th>✓ Hybrid (Streaming + Batch)</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓ Interactive</td>
<td>✓ Interactive</td>
<td>✓ Interactive</td>
<td>✓ Interactive</td>
</tr>
<tr>
<td>✓ Near-Real Time Streaming</td>
<td>✓ Real-Time Streaming</td>
<td>✓ Native Iterative processing</td>
<td>✓ Native Iterative processing</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MapReduce</th>
<th>Direct Acyclic Graphs (DAG) Dataflows</th>
<th>RDD: Resilient Distributed Datasets</th>
<th>Cyclic Dataflows</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>1st Generation (1G)</th>
<th>2nd Generation (2G)</th>
<th>3rd Generation (3G)</th>
<th>4th Generation (4G)</th>
</tr>
</thead>
</table>
5. Apache Flink as the Next Generation of Big Data Analytics

Apache Flink’s **original vision** was getting the best from both worlds: MPP Technology and Hadoop MapReduce Technologies:

- **Declarativity**
- **Query optimization**
- **Efficient parallel in-memory and out-of-core algorithms**

- **Real-Time Streaming**
- **Iterations**
- **Memory Management**
- **Advanced Dataflows**
- **General APIs**

- **Massive scale-out**
- **User Defined Functions**
- **Complex data types**
- **Schema on read**

Draws on concepts from

MPP Database Technology

Draws on concepts from

Hadoop MapReduce Technology
Agenda

1. Capital One
2. Traditional Batch Analytics
3. The Great Paradigm Shift – Real-Time Analytics
4. What are the Drivers?
5. Apache Flink – Next Generation Big Data Analytics Framework
6. Business Use Case: Customer Activity Event Logs
7. Conclusions
6. Business Use Case: Customer Activity Event Logs

1. Customer Activity Log (CAL) Events
2. CAL Analytics Architecture
3. Real-Time Analytics with CAL Data
4. Implementation Details
5. Generic Pattern of Streaming Analytics Architecture
6.1 Business Use Case – Customer Activity Log Events

- Capital One provides many digital platforms for its customers for accomplishing tasks online that were traditionally done manually.
- This is more efficient way to support our customers for their needs and at the same time provides better customer experience.
- It is critical that we make sure our digital platforms are working as intended and detect any issues fast enough to remedy them.
- Customer Activity Logs (CAL) are real-world events of customer activity that is a digital footprint of what a customer is doing.
- CAL events are NOT clickstream data.
- CALs we collect provides valuable data that can be leveraged effectively to achieve the goal of providing a great customer experience.
- CALs standardizes customer activity across applications.
6.2 Architecture of Customer Activity Logs

Systems producing data

Apache Nifi -> Apache Kafka

Apache Flink

Realtime Alerts
Time-Window Aggregates
Transformation & Enrichment

Elasticsearch

Alerts Index
Aggregates Index
Events Index

Consistent long-term storage of stream
Alerts file

Batch processing layer

Kibana Dashboard
6.3 Real-Time Analytics with CAL Data

1. Ability To React to Events in Real-Time – Real-Time Alerts
   - Detecting Fraudulent Devices

2. Real-Time Enrichment
   - Adding information from different sources

3. Real-Time Transformation
   - Flattening nested structure for real-time search and index

4. Real-Time Aggregations
   - Sliding Window based aggregations feeding real-time dashboards

5. Real-Time Index and Search

6.4 Implementation Details

1. Infrastructure setup
2. Real-Time Alerts
3. Real-Time Enrichment
4. Real-Time Transformation
5. Real-Time Aggregations
6. Real-Time Index and Search
6.4.1 Implementation Setup

Infrastructure: Created cluster in AWS
  - Simple 3 Node Cluster

Software
  - Hadoop 2.6.0
  - Flink 0.10-SNAPSHOT as a **YARN** Application
  - ElasticSearch v 1.7.2 Installed on the same cluster
  - Kafka cluster (two node) to feed the real-time stream
  - Kibana v 4.1.2

Data Set: Use Mobile Audit Logging data
  - **Mobile Audit Logging Data** – Sanitized all the sensitive fields with one-way SHA1 hashing
  - Use a file as a source to **generate the streaming data** to feed Kafka.
  - **Live feed** is planned to be done soon
6.4.2 Real-Time Alerts

Alert Conditions JSON

```json
{
  "alerts": [
    {
      "name": "Rule1",
      "type": "condition",
      "lookupfile": "",
      "field": "",
      "lookupNbr": "",
      "condition": "event.EVT_TYPE_CD == '5000023'",
      "message": "Login Error Occurred. Please check"
    }
  ]
}
```

AWS SNS

Can be extended to more options

REST API

Or
6.4.3 Enrichment

```
{
  "EVT_ID": "1",
  "EVT_TS": "2015-08-09 18:00:01.274",
  "EVT_TYPE_CD": "92510"
}
```

```
{
  "EVT_ID": "1",
  "EVT_TS": "2015-08-09 18:00:01.274",
  "EVT_TYPE_CD": "92510",
  "EVENT_DESC": "RetrieveBankLocations"
}
```
6.4.4 Transformations

Transforming JSON array element into individual key value pairs using Jackson serializer Jar.

Example Input:

```
{   "event_id":"1",
    "event_details": [ {
        "detail_key": "user_id",
        "detail_value": "rtmprod-client.kdc.capitalone.com"},
        { 
        "detail_key": "httpStatusCode",
        "detail_value": "409"},
    ] }
```

Output after transformation

```
{   "event_id":"1",
    "user_id": "rtmprod-client.kdc.capitalone.com",
    "httpStatusCode": "409"}
```
6.4.5 Window Aggregates - Time-based Sliding Window

Windows Size = 2 sec
Refresh Interval = 1 sec
6.4.5. Real-Time Index and Search

August 9th, 2015, 08:51:46.464 to August 9th, 2015, 23:31:18.909 — by 10 minutes

Selected Fields
- _source

Available Fields
- _source

Popular
- APPN_SYS_CD
- ACTVY_DISP_CD
- Api-Key
- CRTL_ID
- CRTLNT_ID
- Client-Correlation-ID
- EVENT_DESCRIPTION
- EVT_DETLL_TXT
- EVT_ID
- EVT_STAT_CD
- EVT_TS
- EVT_TYPE_CD

Sample results:

- **USER_ID:** rtmpod-client.kdc.capitalone.com
  - **httpStatusCode:** 409
  - **EVT_ID:** 1,209,171,192
  - **EVT_TS:** August 9th, 2015, 18:47:22.053
  - **EVT_TYPE_CD:** 5,000,028
  - **CRTL_ID:** f55244cc-1612-4f0b544-35af11c01a46
  - **WEB_ANALYTCS_CRTLNT_ID:** 66.87.120.202
  - **APPN_SYS_CD:** EAPI
  - **EVT_STAT_CD:** 201,190
  - **CRET_TS:** August 9th, 2015, 18:47:22.062
  - **EVT_DETLL_TXT:**

- **USER_ID:** rtmpod-client.kdc.capitalone.com
  - **httpStatusCode:** 409
  - **EVT_ID:** 1,209,171,192
  - **EVT_TS:** August 9th, 2015, 18:47:22.053
  - **EVT_TYPE_CD:** 5,000,028
  - **CRTL_ID:** f55244cc-1612-4f0b544-35af11c01a46
  - **WEB_ANALYTCS_CRTLNT_ID:** 66.87.120.202
  - **APPN_SYS_CD:** EAPI
  - **EVT_STAT_CD:** 201,190
  - **CRET_TS:** August 9th, 2015, 18:47:22.062
  - **EVT_DETLL_TXT:**
6.5 Generic Pattern Supports A Class of Use Cases

Event Producers
- Apps
- Devices
- Sensors

Event Collector
- Flume
- SpringXD
- Logstash
- Nifi
- Fluentd

Event Broker
- Kafka
- RabbitMQ
- JMS

Event Processor
- Flink
- Spark
- Storm
- Samza

Indexer
- ElasticSearch
- Solr
- Cassandra
- NoSQL DB

Dashboard
- Kibana
- D3
- Custom GUI

Dynamic Models

RealTime Actions

Notifications

Dynamic Models

RealTime Actions

Notifications
6.6 The Analytics Spectrum – Batch & Real-Time

Input Stream -> Depth of Analysis -> Latency

- Quick Aggregations/Alerts
- Intermediate
- Deep Learning

Insights
Agenda

1. Capital One
2. Traditional Batch Analytics
3. The Great Paradigm Shift – Real-Time Analytics
4. What are the Drivers?
5. Apache Flink – Next Generation Big Data Analytics Framework
6. Business Use Case: Customer Activity Event Logs
7. Conclusions
6. Conclusions & Key Takeaways

- Traditional Batch Analytics has long intervals from data to insights and insights to action (CSAD Cycles)
- **Business, Technological and Social Drivers** and demanding time to insights and action in seconds, not days
- New Streaming Technologies such as **Apache Flink** enabling Enterprises to react to events in real-time **as-they-happen**
- Future Competitiveness of Business rests on the ability to capture, move, and process large amounts of data in real-time.
- **Paradigm shift towards Fast Data** is happening across enterprises. It is not an option, it is a must for any business.
- There is still **room for batch analytics**, but lot of todays workloads will move to Streaming Real-Time Analytics and continuous ETL.
Thank You!

**Capital One** is hiring for multiple locations

http://jobs.capitalone.com and search on: #ilovedata.

**Stay In Touch**

spalthepu@gmail.com  
@SriniPalthebu  
https://www.linkedin.com/in/srinipalthebu