Open Source Data Management for Car-to-Cloud

Michael Ger
Managing Director, Manufacturing & Automotive
Cloudera
CLOUDERA AT A GLANCE

One stop shop for Big Data management for analytics

Unified open source architecture

Hybrid and multi-cloud
CLOUDERA ENTERPRISE DATA WAREHOUSE

Any Data. Anywhere. From Edge to AI

Traditional Data Warehouse Optimization
- BI Reporting
- Data Warehouse Offload
- ETL Optimization

Operations & Events Data Warehouse
- IOT Data (Sensors, Logs, etc.)
- Connected Vehicles
- Connected Factories (Industry 4.0)
- Etc.

Research & Discovery Data Warehouse
- Unstructured Data (Images, Video, etc.)
- Intelligent Search
- Autonomous Vehicle Research
10 of the top 10 Automotive OEMs
Trust Cloudera for Big Data-Enabled Digital Transformations
DATA TRENDS

Connected Vehicle Data

10X by 2020*

* Source: Cowen and Company, Gartner

Autonomous Vehicle Data

TERabytes per vehicle, per day

Industrial Internet Data

2X faster than any other data source

* Source: Wikibon

Real-Time Data

1.5X faster than traditional data

* Source: IDC
<table>
<thead>
<tr>
<th>Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Edge</td>
</tr>
<tr>
<td>Managing the Edge</td>
</tr>
<tr>
<td>Enterprise Flow/Stream Analytics</td>
</tr>
<tr>
<td>Edge to AI</td>
</tr>
<tr>
<td>Connected Communities</td>
</tr>
</tbody>
</table>
CAR TO CLOUD – DATA MANAGEMENT ON 5 LEVELS

<table>
<thead>
<tr>
<th>Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>THE EDGE</td>
</tr>
<tr>
<td>MANAGING THE EDGE</td>
</tr>
<tr>
<td>ENTERPRISE FLOW/STREAM ANALYTICS</td>
</tr>
<tr>
<td>EDGE TO AI</td>
</tr>
<tr>
<td>CONNECTED COMMUNITIES</td>
</tr>
</tbody>
</table>
EDGE DATA COLLECTION - MINIFI AGENTS

- Extremely small footprint
- Java and C++ agents
  - JRHEL/CentOS, Debian/Ubuntu, Android*
- Key data management functions
  - Filtering, buffering, guaranteed delivery, prioritized queuing
- Secure
  - Encryption, certificate-based authentication
  - Data tagging and provenance
- Execute Machine Learning (ML) models including Tensorflow
GOING SMALLER – FROM NIFI TO MINIFI
Supporting the Need for Smaller Footprints on the Edge

<table>
<thead>
<tr>
<th>Application</th>
<th>Memory Footprint</th>
<th>Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>MiNiFi Agent (Java)</td>
<td>10s to 100s of MBs</td>
<td>• Feature parity and reuse of core NiFi libraries</td>
</tr>
<tr>
<td>MiNiFi Agent (C++)</td>
<td>100s kBs to MBs</td>
<td>• Write once**, run anywhere.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Adaptable to lower level interfaces</td>
</tr>
<tr>
<td>MiNiFi Library</td>
<td>10s kBs</td>
<td>• Write n-many times, embed, run anywhere</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Language libraries to support tagging, FlowFile format, Site to Site protocol and provenance generation without a full processing framework</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Customizable to environments without disk storage or where threading is prohibitive</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Language SDKs, Mobile Platforms</td>
</tr>
</tbody>
</table>
MINIFI IN CONNECTED CAR

Processing/Synthesis
- Transmit
- Execute
- Filter
- Prioritize
- Route

Comprehension
- Parse CAN
- Parse Ethernet
- Parse LIN
- Parse <>

Collection
- Listen CAN
- Listen Ethernet
- Listen LIN
- Listen <>

Gateway
- CAN Bus
  - MCU
  - MCU
  - MCU

Ethernet/ Ethernet AVB

Local Interconnect Network

Yet to be established protocol

MiNiFi Agent

MiNiFi Library

MiNiFi IN CONNECTED CAR
<table>
<thead>
<tr>
<th>Level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>THE EDGE</td>
<td></td>
</tr>
<tr>
<td>MANAGING THE EDGE</td>
<td></td>
</tr>
<tr>
<td>ENTERPRISE FLOW/STREAM ANALYTICS</td>
<td></td>
</tr>
<tr>
<td>EDGE TO AI</td>
<td></td>
</tr>
<tr>
<td>CONNECTED COMMUNITIES</td>
<td></td>
</tr>
</tbody>
</table>
Edge Agent Management
- Central management of agents
- Collect data from edge device
- Push intelligence to edge

Data Collection and Processing at the Edge
- Small, lightweight footprint
EDGE FLOW MANAGER

• Edge Management Hub
• Graphical user interface to develop and deploy flows to edge
• Monitor thousands of edge agents
• Deploy updated Machine Learning (ML) models to edge agents
• Integration with NiFi Registry
EFM will version the flow into the NiFi Registry and expose an endpoint for all agents to automatically download and deploy the flow. Powerful Edge Design/Deploy Pattern!
<table>
<thead>
<tr>
<th>Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>THE EDGE</td>
</tr>
<tr>
<td>MANAGING THE EDGE</td>
</tr>
<tr>
<td>ENTERPRISE FLOW/STREAM ANALYTICS</td>
</tr>
<tr>
<td>EDGE TO AI</td>
</tr>
<tr>
<td>CONNECTED COMMUNITIES</td>
</tr>
</tbody>
</table>
FLOW MANAGEMENT VIA NIFI

• Based on Apache NiFi
• Highly configurable flow creation
• Web-based user interface
• 300+ prebuilt processors
• Secure - fine grained encryption
• Data provenance
• Guaranteed delivery
  - Buffer data during system interruptions
• Designed for extensibility
• NiFi Registry
WHAT CAN BE DONE WITH NIFI PROCESSORS?
300+ Pre-Built Processors

- **Ingestion**: connectors to read/write data from/to several data sources
  - **Protocols**: HTTP (S), AMQP, MQTT, UDP, TCP, CEF, JMS, (S) FTP, AWSIoT, Raw Socket Protocol
  - **Brokers**: Kafka, JMS, AMQP, MQTT etc.
  - **Databases**: JDBC, MongoDB, HBase, Cassandra etc.

- **Extraction** (XML, JSON, Regex, Grok etc.)

- **Transformation**
  - Format conversion (JSON to Avro, CSV to ORC etc.)
  - Compression/decompression, Merge, Split, encryption etc.

- **Data enrichment**
  - Attribute, content, rules etc.

- **Routing**
  - Priority, dynamic/static, based on content or metadata etc.
300+ PROCESSORS FOR DEEPER ECOSYSTEM INTEGRATION

FTP  SFTP  HL7  UDP  XML

HTTP  WebSocket  Email  HTML  Image  Syslog  AMQP

Hash  Encrypt  GeoEnrich
Merge  Tail  Scan
Extract  Evaluate  Replace
Duplicate  Execute  Translate
Split  Fetch  Convert

Parse Records  Convert Records
Route Text  Distribute Load
Route Content  Generate Table Fetch
Route Context  Jolt Transform JSON
Control Rate  Prioritized Delivery

All Apache project logos are trademarks of the ASF and the respective projects.
EXAMPLE - LAS VEGAS CONNECTED VEHICLE PILOT

City Fleet Vehicle In Operation

Monitor Vehicle Conditions

Approaching Active Bus Stop?

Issue Alert to In-Vehicle Display Unit
CITY OF LAS VEGAS – DATA AND SYSTEMS

Vehicles

Speed and Location

Vehicle Communications

RVI Server

Vehicle Messages

Vehicle Alerts

Application Logic & Message Routing

nifi

Vehicle Alerts

Real-Time City Data Sources

(Bus Locations)

(Bus Stops)

Static City Data Sources

(Speed Limits)

(Pedestrian Zones)

Long Term Storage

Reference Data

Vehicle Messages

Static Data

cloudera
Data Ingestion

1. Receive Vehicle Data (ListenHTTP)
2. Fork Data
3a. Extract GPS Data
3b. Extract Vehicle Speed Data
4. Reformat for each use case
5. Output Data to Use Cases
6. Store in Hadoop

- To UC1
- To UC2
- To UC3
1. Receive UC3 RVI Data (Location)

2. Build location based Query

3. Query if vehicle approaching bus stop location

4. Determine if bus stop is ACTIVE

5. To Active Bus Stop Warning
CAR TO CLOUD – DATA MANAGEMENT ON 5 LEVELS

THE EDGE

MANAGING THE EDGE

ENTERPRISE FLOW/STREAM ANALYTICS

EDGE TO AI

CONNECTED COMMUNITIES
CONNECTED VEHICLE ANALYTICS LIFECYCLE
Example: Vehicle Predictive Maintenance

1. INGEST
2. STORE
3. PROCESS
4. ANALYZE
5. LEARN
6. DEPLOY
7. ACT

REAL-TIME DATA SOURCES
Connected Vehicle

ENTERPRISE TRANSACTION DATA
Dealer Service, Warranty, MFG, Etc.

ENTERPRISE DATA LAKE

REAL-TIME ACTION

Model Inputs
- Historic vehicle CAN Bus values
- Historic maintenance records
- Historic usage characteristics
- Historic failures

- Machine Learning
- Train Scoring Models
- Data Discovery
- Business Intelligence
SOLUTION COMPONENT DETAIL

REAL-TIME DATA SOURCES
Connected Vehicle

REAL-TIME ACTION
ACT

Model Inputs
- Historic vehicle CAN Bus values
- Historic maintenance records
- Historic usage characteristics
- Historic failures

STORE AND PROCESS

ENTERPRISE DATA LAKE

INGEST AND DEPLOY

DEPLOY

LEARN
CDSW
Spark ML

ANALYZE
druid

INGEST

ENTERPRISE TRANSACTION DATA
Dealer Service, Warranty, MFG, Etc.
CLOUDERA FOR CONNECTED VEHICLE INNOVATION

REAL-TIME DATA SOURCES

Connected Vehicle

INGEST

DEPLOY

STORE AND PROCESS

ENTERPRISE DATA LAKE

1. INGEST
2. STORE AND PROCESS
3. ENTERPRISE DATA LAKE
4. ANALYZE
5. LEARN
6. DEPLOY
7. REAL-TIME ACTION

Model Inputs
- Historic vehicle CAN Bus values
- Historic maintenance records
- Historic usage characteristics
- Historic failures

Model Outputs
- Data Discovery
- Business Intelligence

Data Discovery and Business Intelligence

1. CDF DATA IN MOTION
2. CDH DATA AT REST
3. HDP DATA AT REST

Dealer Service, Warranty, MFG, Etc.
CAR TO CLOUD – DATA MANAGEMENT ON 5 LEVELS

- **THE EDGE**
- **MANAGING THE EDGE**
- **ENTERPRISE FLOW/STREAM ANALYTICS**
- **EDGE TO AI**
- **CONNECTED COMMUNITIES**
CONNECTED VEHICLE IS LEADING THE TRANSFORMATION
Monetizing the Connected Vehicle

- Data-driven policy decision-making without data liability
- Dynamic interactions with close proximity target audiences
- Dynamic transportation & infrastructure mgt. and establishing merchant community hub
- Real-time service for lifetime relationship with Customer
- Real-time grid capacity management based on vehicle movements
- Improve consumer quality of life via car/home/work integration (i.e. security, climate, home supply mgt.)
ESTABLISHING TRUST – WHAT’S REQUIRED?

Data Providers

- Control who can see what data
- Control what data can be combined

End-Users

ASK QUESTION
- Secured views of data and insights

DELIVER INSIGHT

Connected Communities

✔ Data access across systems

✔ Controlling who sees what data

✔ Data audit across the ecosystem

- Each data provider shares specific data assets and defines specific data sharing rules

- Each data provider precisely defines which end-users can see their data throughout the data lifecycle

- Each data provider can track data usage across the data lifecycle, including the data’s origin, where it is used, and how it changes over time
### HOW IT WORKS.....

#### CLOUDERA DATAPLANE SERVICE
Centralized Management of Core Functions

<table>
<thead>
<tr>
<th>CLUSTER MANAGEMENT</th>
<th>USER &amp; SECURITY MANAGEMENT</th>
<th>DATA LIFECYCLE MANAGEMENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>OEM 1</td>
<td>Role 1</td>
<td>GATE</td>
</tr>
<tr>
<td>OEM 2</td>
<td>Role 2</td>
<td>Ingest</td>
</tr>
<tr>
<td>Government</td>
<td>Role 3</td>
<td>Transform</td>
</tr>
<tr>
<td>Researcher</td>
<td>Role 4</td>
<td>Edit</td>
</tr>
<tr>
<td>Retailer</td>
<td></td>
<td>Export</td>
</tr>
<tr>
<td>Home</td>
<td></td>
<td>GATE</td>
</tr>
</tbody>
</table>

- **Cluster registration and visibility** across on-premise, Cloud and Hybrid deployments
- **User authentication** synched with enterprise identity systems
- **Centralize data security roles, privileges and policies** across clusters
- **Track Data Lineage** across data lifecycle
- **Track Data Audit** across data lifecycle
- **Complimentary with BlockChain**

**Connected Vehicle Community**

**Data Providers**

**End-Users**
- Ask Question
- Deliver Insight

**Ingest**
**Transform**
**Edit**
**Export**

**Blockchain GATE**
Single View of All Data Sources
Asset Collections
## Asset Collection Details

### Vehicle Accident Data

Data relating to vehicle accidents

#### Tables

<table>
<thead>
<tr>
<th>Source</th>
<th>Name</th>
<th>Database Name</th>
<th>Owner</th>
<th>Created Time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Emergency Contacts</td>
<td>OEM_owner_portal</td>
<td>Automaker Admin</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Vehicle Events</td>
<td>OEM_vehicle_events</td>
<td>Automaker Admin</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Emergency Incidents</td>
<td>OEM_emergency_details</td>
<td>Automaker Admin</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Emergency Incident Analysis</td>
<td>OEM_emergency_aggregation</td>
<td>Automaker Admin</td>
<td></td>
</tr>
</tbody>
</table>
Tagging of Sensitive Information
Adding Security Policies

Table and Field Level Controls

<table>
<thead>
<tr>
<th>Policy ID</th>
<th>Policy Name</th>
<th>Status</th>
<th>Audit Logging</th>
<th>Group</th>
<th>Users</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>access.CITY_emergency_analysis_table</td>
<td>ENABLED</td>
<td>ENABLED</td>
<td>public</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>mask: name</td>
<td>ENABLED</td>
<td>ENABLED</td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>mask: ssn show first 4</td>
<td>ENABLED</td>
<td>ENABLED</td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>mask: telephone</td>
<td>ENABLED</td>
<td>ENABLED</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Data Lineage (For Table)

Customer Emergency Contact Information

Vehicle Event Data

Accident Detail (For Use by City Emergency Dispatch)

Aggregated Detail (For Use by Various Research Entities)
### Audit Trail

#### Event Time

<table>
<thead>
<tr>
<th>Policy ID</th>
<th>Event Time</th>
<th>User</th>
<th>Resource Type</th>
<th>Access Type</th>
<th>Result</th>
<th>Access Type</th>
<th>Client IP</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>07/17/2018 16:14:55 GMT</td>
<td>katu_trv</td>
<td>@column</td>
<td>SELECT</td>
<td>ALLOWED</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>07/17/2018 16:14:55 GMT</td>
<td>katu_trv</td>
<td>@column</td>
<td>SELECT</td>
<td>ALLOWED</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>07/17/2018 16:08:37 GMT</td>
<td>katu_trv</td>
<td>@column</td>
<td>SELECT</td>
<td>ALLOWED</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>07/17/2018 16:06:44 GMT</td>
<td>mark_zhao</td>
<td>@column</td>
<td>SELECT</td>
<td>ALLOWED</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>07/17/2018 16:05:26 GMT</td>
<td>diana_psr</td>
<td>@column</td>
<td>SELECT</td>
<td>DENIED</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>07/17/2018 16:02:46 GMT</td>
<td>elf_user</td>
<td>@column</td>
<td>SELECT</td>
<td>DENIED</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>07/17/2018 16:02:05 GMT</td>
<td>katu_trv</td>
<td>@column</td>
<td>SELECT</td>
<td>ALLOWED</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>07/17/2018 16:02:05 GMT</td>
<td>katu_trv</td>
<td>@column</td>
<td>SELECT</td>
<td>ALLOWED</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>07/12/2018 14:14:36 GMT</td>
<td>jemmy_contractor</td>
<td>@column</td>
<td>SELECT</td>
<td>ALLOWED</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>07/12/2018 14:14:36 GMT</td>
<td>jemmy_contractor</td>
<td>@column</td>
<td>SELECT</td>
<td>ALLOWED</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>07/12/2018 14:13:52 GMT</td>
<td>diana_psr</td>
<td>@column</td>
<td>SELECT</td>
<td>DENIED</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>07/12/2018 14:10:36 GMT</td>
<td>jemmy_contractor</td>
<td>@column</td>
<td>SELECT</td>
<td>DENIED</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>07/12/2018 14:07:29 GMT</td>
<td>ivsna_arh</td>
<td>@column</td>
<td>SELECT</td>
<td>DENIED</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>07/12/2018 14:05:54 GMT</td>
<td>joe_analyst</td>
<td>@column</td>
<td>MASK_NIL</td>
<td>ALLOWED</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>07/12/2018 14:35:54 GMT</td>
<td>joe_analyst</td>
<td>@column</td>
<td>CUSTOM</td>
<td>ALLOWED</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>07/12/2018 14:35:54 GMT</td>
<td>joe_analyst</td>
<td>@column</td>
<td>CUSTOM</td>
<td>ALLOWED</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>07/12/2018 14:35:54 GMT</td>
<td>joe_analyst</td>
<td>@column</td>
<td>CUSTOM</td>
<td>ALLOWED</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>07/12/2018 14:35:54 GMT</td>
<td>joe_analyst</td>
<td>@column</td>
<td>CUSTOM</td>
<td>ALLOWED</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>07/12/2018 14:35:54 GMT</td>
<td>joe_analyst</td>
<td>@column</td>
<td>CUSTOM</td>
<td>ALLOWED</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
CONCLUSION

• Connected Vehicle Data is Key to Industry Transformation

• Evolving Intelligence at Edge and Connected Community Use Cases

• Big Data Management Challenges on Multiple Levels

• Open Source Data Management Innovation Can Help!
THANK YOU