



Next-gen Data Flow Platform for the Enterprise

Santosh Bardwaj
Vice President, Advanced Analytics

The opinions expressed in this presentation are those of the presenters,
in their individual capacities, and not necessarily those of Discover.

Agenda

1

**Discover's
next-gen data
ingestion
platform
built on NiFi**

2

**What it
takes to
build an
enterprise-
ready
platform**

3

**Challenges
and how we
overcame
them**

4

**Next steps
with the
platform**

Discover is a leading U.S. direct bank & payments partner

\$60Bn in Credit Card Receivables

1 in 4 Households¹



Leading Cash
Rewards

5%
CASHBACK
BONUS



DISCOVER
Deposits & Lending

\$37Bn Consumer Deposits

\$9Bn Private Student Loans

\$7Bn Personal Loans



- \$183Bn Payment Services Volume
- 185+ Countries/Territories

Note(s)

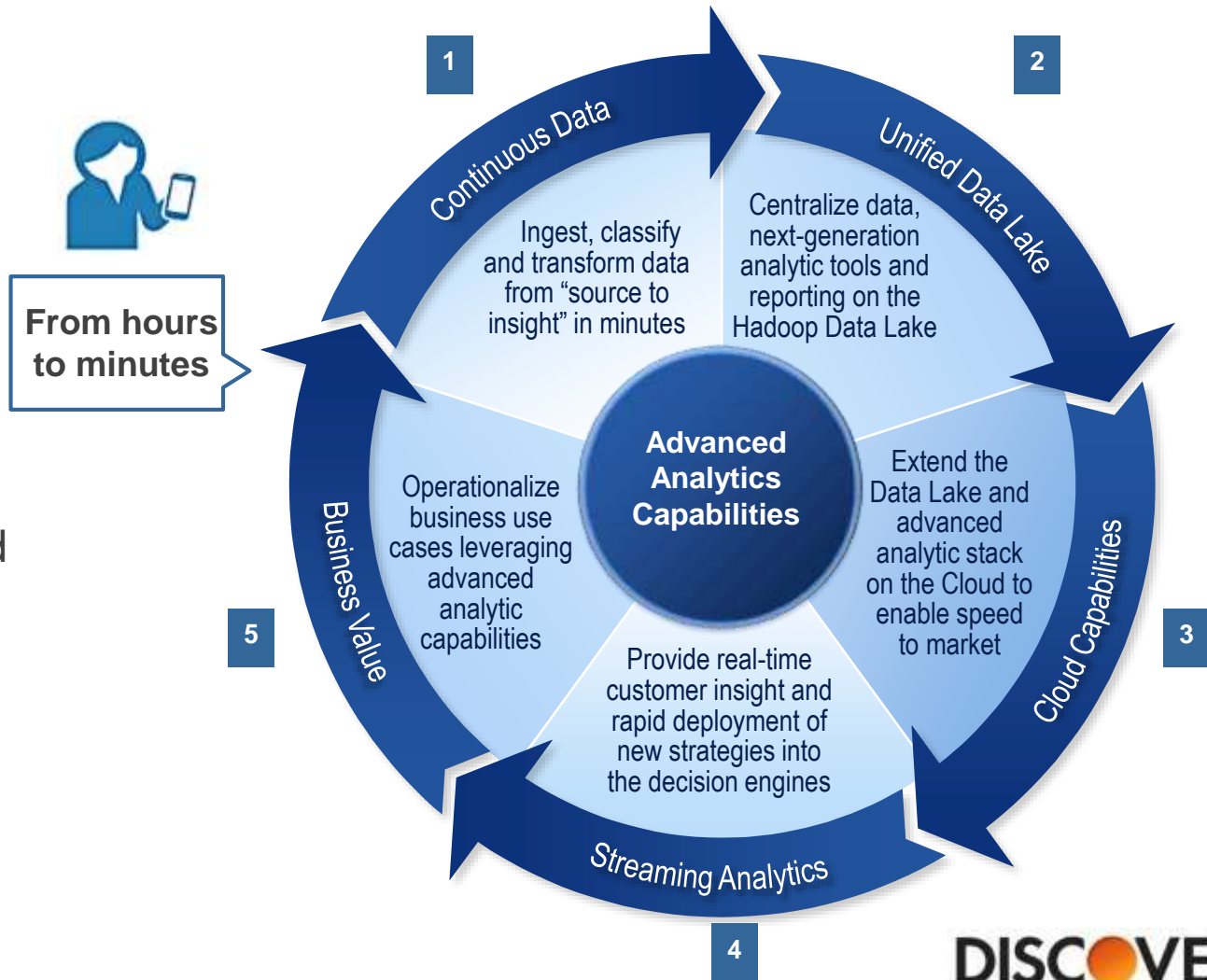
Balances as of March 31, 2017; volume based on the trailing four quarters ending 1Q17; direct-to-consumer deposits includes affinity deposits

1. TNS' Consumer Payment Strategies Study

DISCOVER

Advancing our data-analytic capabilities

- Built around a foundation of a continuous data pipeline and hybrid data-analytic lake



Unified data ingestion platform

- Ingest data from source systems
- Push to the Enterprise Data Lake
- Governed process leveraging common-reusable templates

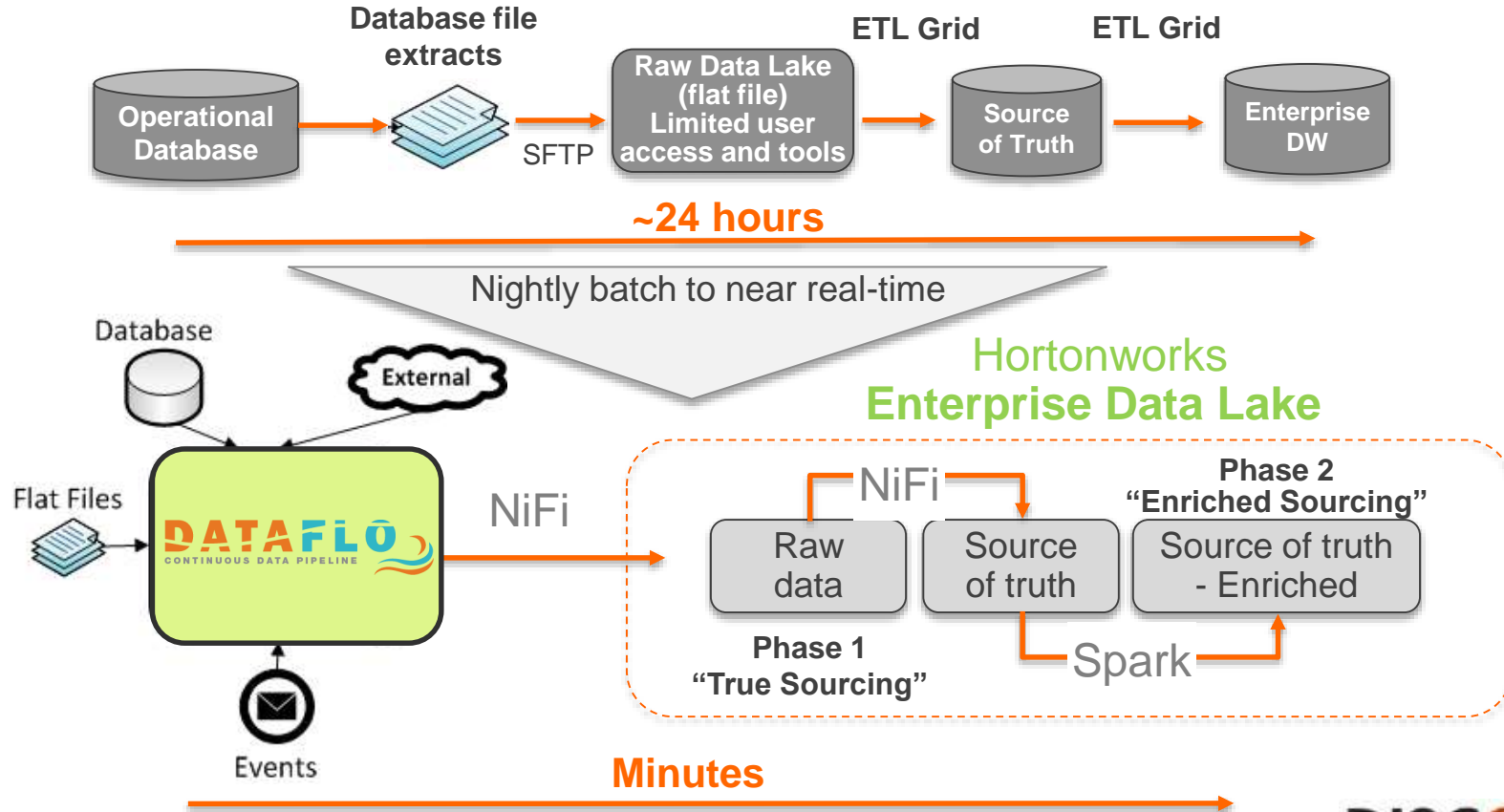
What is NiFi?

- Enables automated data flow management
- Acquires data from producers
- Delivers to consumers while orchestrating the flow

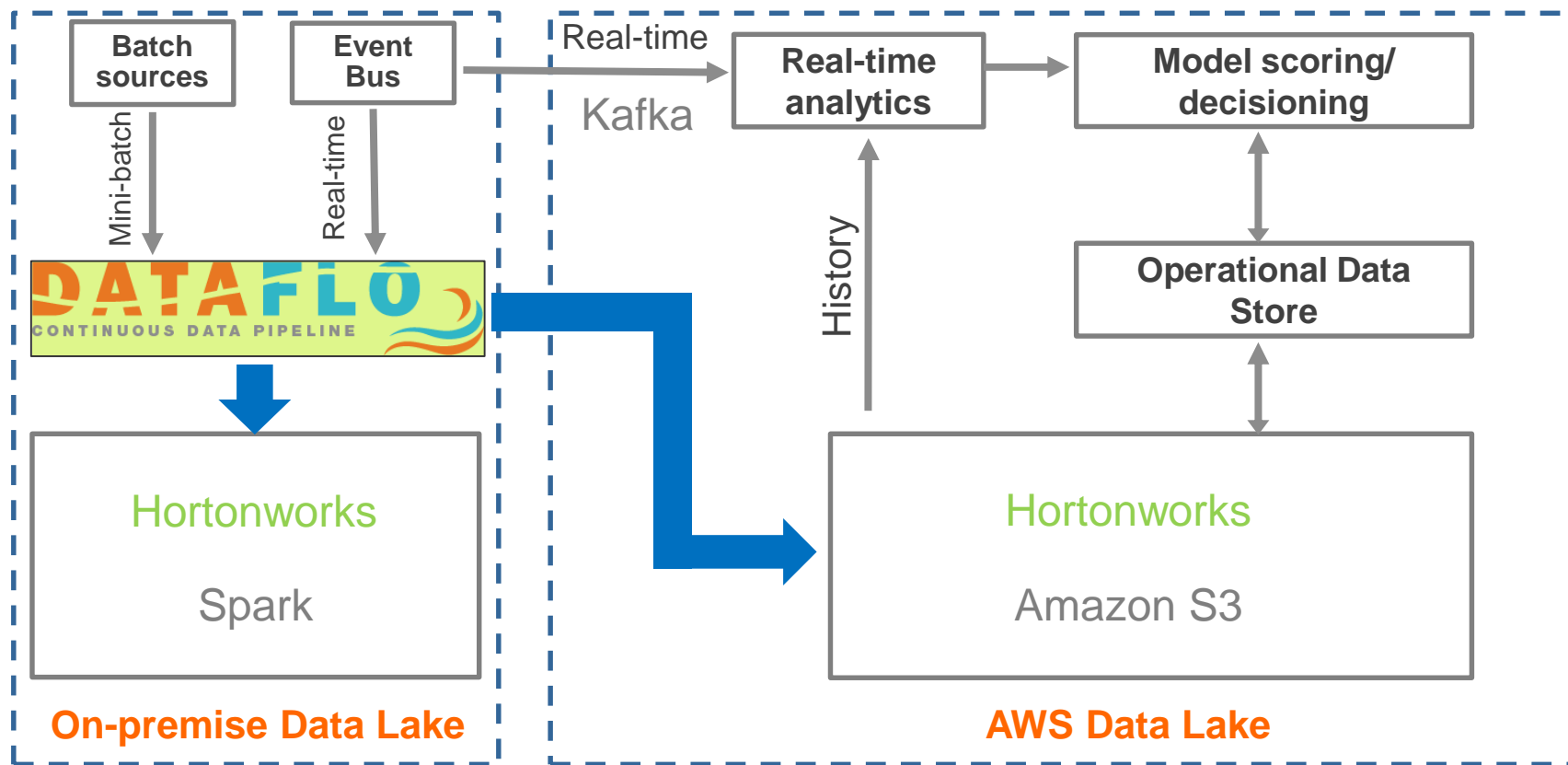
Why we chose NiFi to build our data ingestion platform

- ☐ Scalable and Customizable
- ☐ Provenance
- ☐ Promotes reuse
- ☐ Secure
- ☐ User Interface (drag & drop)

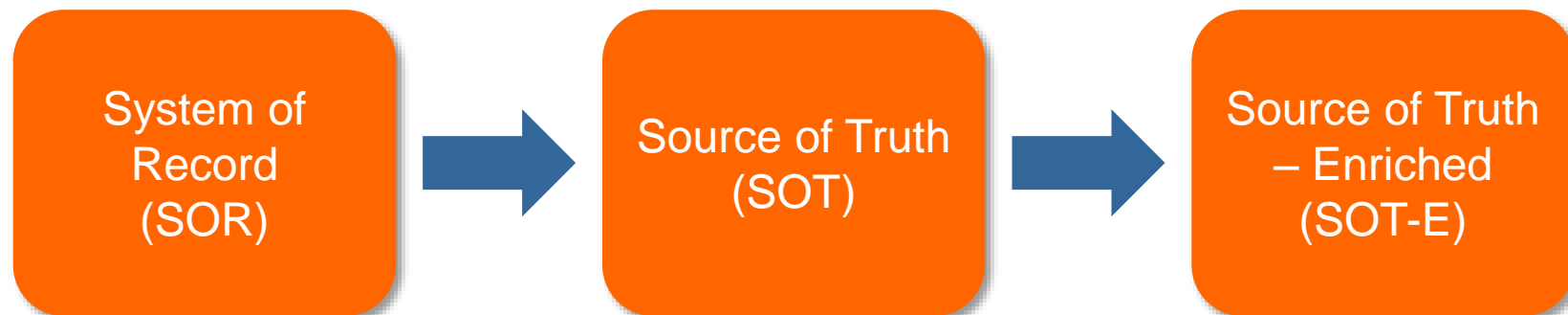
The next-gen **DATAFLO** CONTINUOUS DATA PIPELINE platform built on NiFi and Spark is designed to streamline our data pipeline into a near real-time paradigm



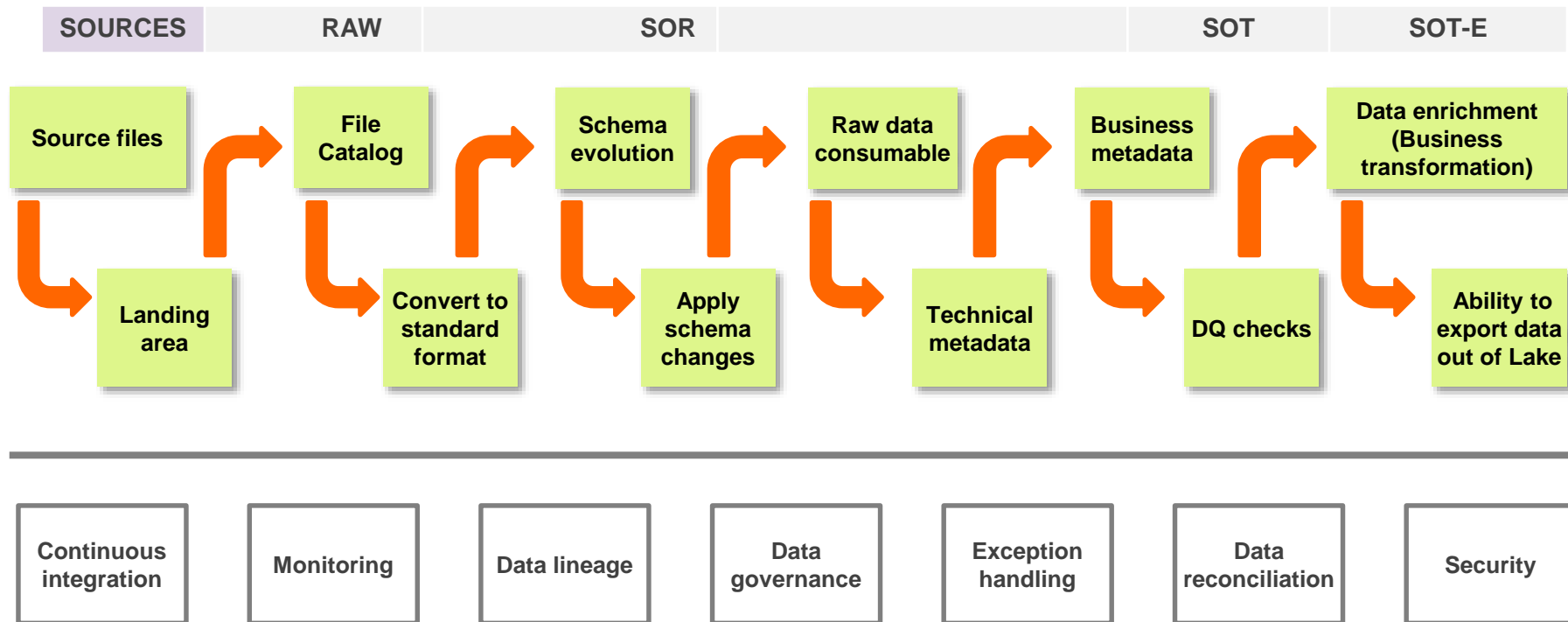
We are also extending the capability of **DATAFLO** CONTINUOUS DATA PIPELINE into the cloud



Data Flow Categorization within the Hadoop Data Lake



Detail flow and foundational components



Ingesting complex data - How complex?

Format of files will vary, some are easy to consume, others hard

Example: Records with Dynamic arrays/vectors of primitives or strings

Schema: First Name, Last Name, Array_size of Sibling_Name[], Sibling_Name[0-N], City

Data:

```
John, Doe, 2, Susie, Chris, Chicago
Mary, Johnston, 3, Ashley, Tom, Mike, Atlanta
Frank, Smith, 1, Ralph, Toronto
```

Example: Records with an array of Struct data types

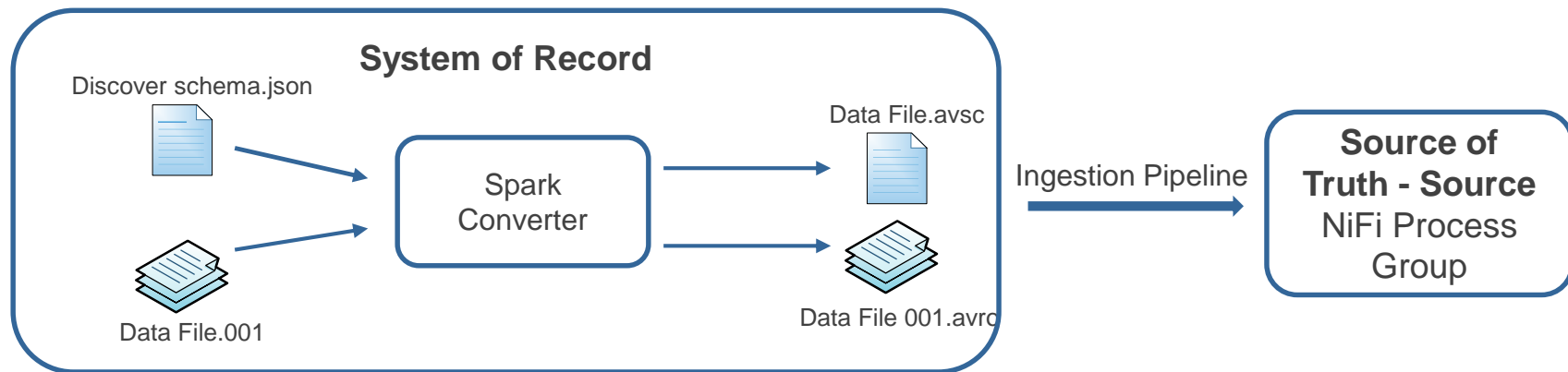
Schema: First Name, Array_size of CompanyStruct[], CompanyStruct.Name, CompanyStruct.City, CompanyStruct.YearsWorked, Age

Data:

```
John, 1, Discover, Chicago, 3, 44
Mary, 3, Sales Unlimited, Dallas, 2, Auditors R' Us, Atlanta, 5, Discover, Chicago 4, 35
```



Our solution – A custom NiFi processor to handle complex data types



Continuous improvement of real-time data ingestion using NiFi

NiFi Ingestion Flow Version I

24 hours

Source : Flat File

 GetFile			
In	0 (0 bytes)	5 min	
Read/Write	0 bytes / 0 bytes	5 min	
Out	0 (0 bytes)	5 min	
Tasks/Time	0 / 00:00:00.000	5 min	

 ConvertCSVToAvro			
In	0 (0 bytes)	5 min	
Read/Write	0 bytes / 0 bytes	5 min	
Out	0 (0 bytes)	5 min	
Tasks/Time	0 / 00:00:00.000	5 min	

 MergeContent			
In	0 (0 bytes)	5 min	
Read/Write	0 bytes / 0 bytes	5 min	
Out	0 (0 bytes)	5 min	
Tasks/Time	0 / 00:00:00.000	5 min	


Destination: Hadoop


 PutHDFS			
In	0 (0 bytes)	5 min	
Read/Write	0 bytes / 0 bytes	5 min	
Out	0 (0 bytes)	5 min	
Tasks/Time	0 / 00:00:00.000	5 min	


NiFi Ingestion Flow Version II

Complex logic, limited scale

Source : Event Bus


 ConsumeJMS			
In	0 (0 bytes)	5 min	
Read/Write	0 bytes / 0 bytes	5 min	
Out	0 (0 bytes)	5 min	
Tasks/Time	0 / 00:00:00.000	5 min	

 ExtractText			
In	0 (0 bytes)	5 min	
Read/Write	0 bytes / 0 bytes	5 min	
Out	0 (0 bytes)	5 min	
Tasks/Time	0 / 00:00:00.000	5 min	

 ReplaceText			
In	0 (0 bytes)	5 min	
Read/Write	0 bytes / 0 bytes	5 min	
Out	0 (0 bytes)	5 min	
Tasks/Time	0 / 00:00:00.000	5 min	

 UpdateAttribute			
In	0 (0 bytes)	5 min	
Read/Write	0 bytes / 0 bytes	5 min	
Out	0 (0 bytes)	5 min	
Tasks/Time	0 / 00:00:00.000	5 min	

Destination: Hadoop


 PutHDFS			
In	0 (0 bytes)	5 min	
Read/Write	0 bytes / 0 bytes	5 min	
Out	0 (0 bytes)	5 min	
Tasks/Time	0 / 00:00:00.000	5 min	

NiFi Ingestion Flow Version III

Seconds

Custom NiFi processor developed in-house, reusable and scalable


Source : Event Bus

 ConsumeJMS			
In	0 (0 bytes)	5 min	
Read/Write	0 bytes / 0 bytes	5 min	
Out	0 (0 bytes)	5 min	
Tasks/Time	0 / 00:00:00.000	5 min	

 ConvertKeyValuePayloadtoCSV			
In	0 (0 bytes)	5 min	
Read/Write	0 bytes / 0 bytes	5 min	
Out	0 (0 bytes)	5 min	
Tasks/Time	0 / 00:00:00.000	5 min	

 ConvertAvroToORC			
In	0 (0 bytes)	5 min	
Read/Write	0 bytes / 0 bytes	5 min	
Out	0 (0 bytes)	5 min	
Tasks/Time	0 / 00:00:00.000	5 min	

Destination: Hadoop

 PutHDFS			
In	0 (0 bytes)	5 min	
Read/Write	0 bytes / 0 bytes	5 min	
Out	0 (0 bytes)	5 min	
Tasks/Time	0 / 00:00:00.000	5 min	

ETL on Hadoop progression



Data enrichment from SOR to SOT (~600 jobs)

Version I

Traditional
ETL tool

Version II

ETL on
HiveQL

Version III

ETL on Spark
(hand-coded)

Coming soon

Automated
(flow-based)
ETL on Spark

Run time: ~18 hours

~8 hours

~1 hour



Upcoming enhancements to our data pipeline



Integrating
data
quality,
catalog into
NiFi flow

Custom
processors
to parse
complex
data
structures

Enterprise
scale ETL
on Hadoop
using
Spark

Self-
service
data
pipelines

Integrating
batch and
real-time
data
pipelines



Hiring Data Engineers

Q & A