Next-gen Data Flow Platform for the Enterprise

Santosh Bardwaj
Vice President, Advanced Analytics

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 Households[1]

Leading Cash Rewards

5% CASHBACK BONUS

DISCOVER

Deposits & Lending

$37Bn Consumer Deposits

$9Bn Private Student Loans

$7Bn Personal Loans

Diners Club INTERNATIONAL

DISCOVER

pulse A DISCOVER COMPANY

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

Note(s)

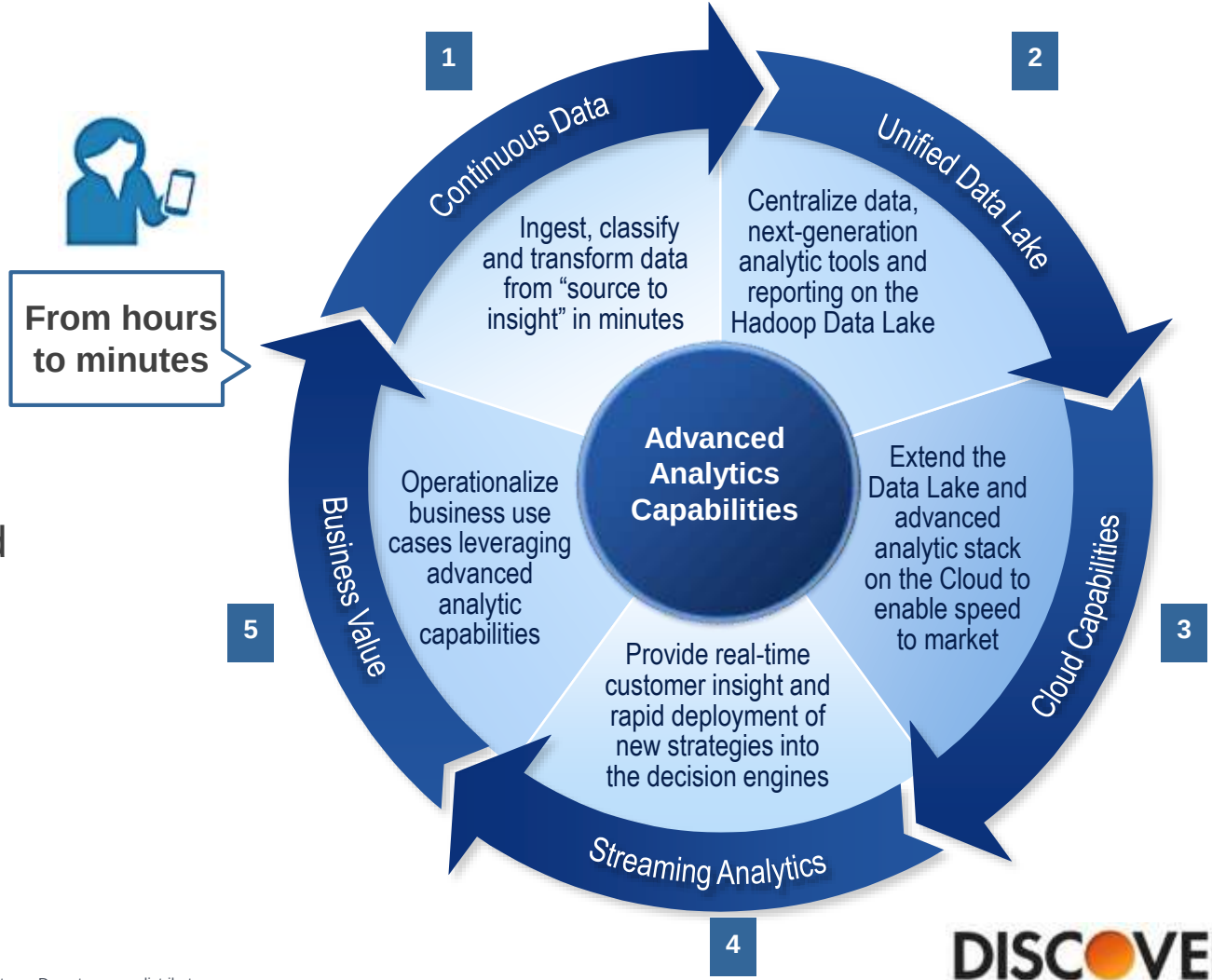
Balances as of March 31, 2017; volume based on the trailing four quarters ending 1Q17; direct-to-consumer deposits includes affinity deposits

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 built on NiFi

DATAFLO
CONTINUOUS DATA PIPELINE

Unified data ingestion platform

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

What is NiFi?

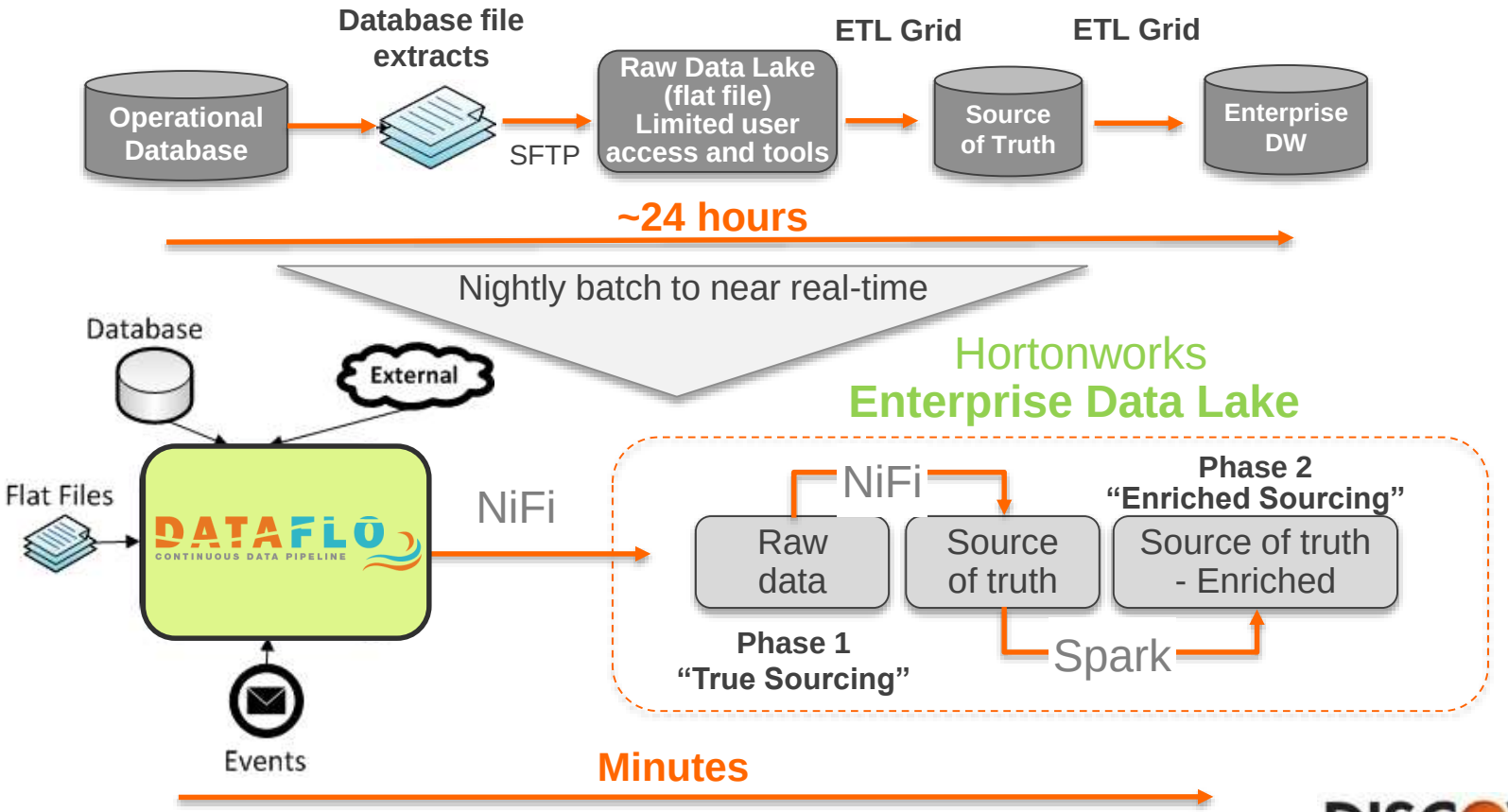
- Enables automated data flow management
- Acquires data from producers
- Delivers to consumers while orchestrating the flow

Why we chose NiFi to build our data ingestion platform

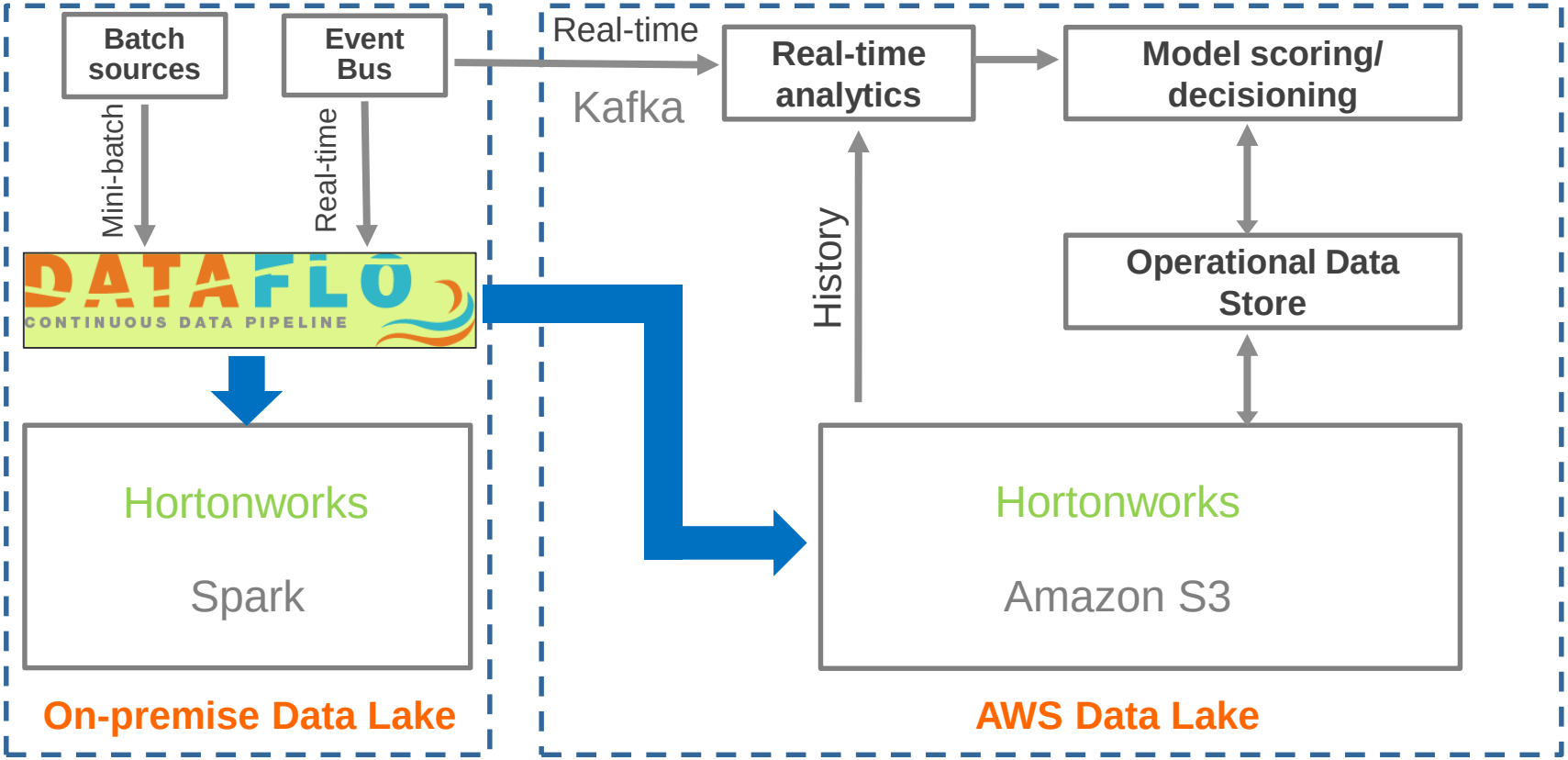
- ☐ Scalable and Customizable
- ☐ Provenance
- ☐ Promotes reuse
- ☐ Secure
- ☐ User Interface (drag & drop)

DISCOVER

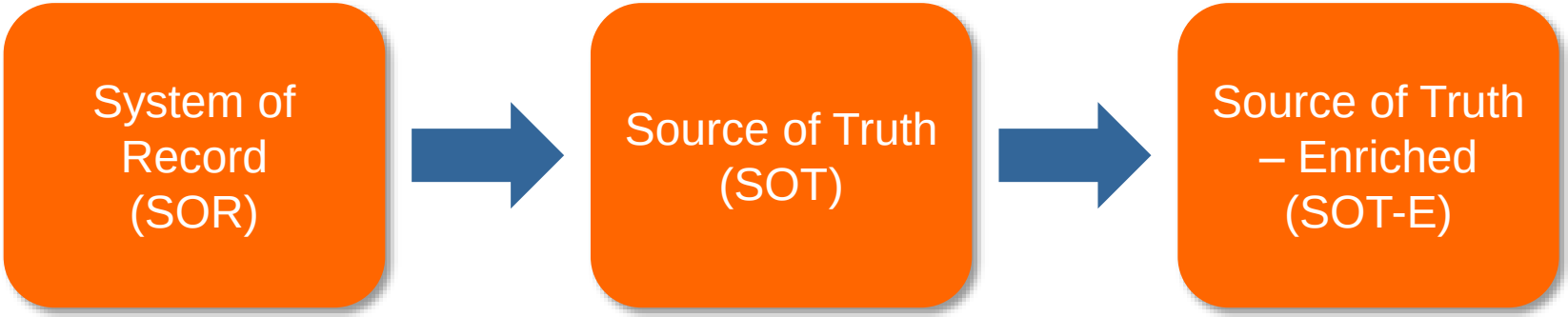
The next-gen DATAFLO platform built on NiFi and Spark is designed to streamline our data pipeline into a near real-time paradigm



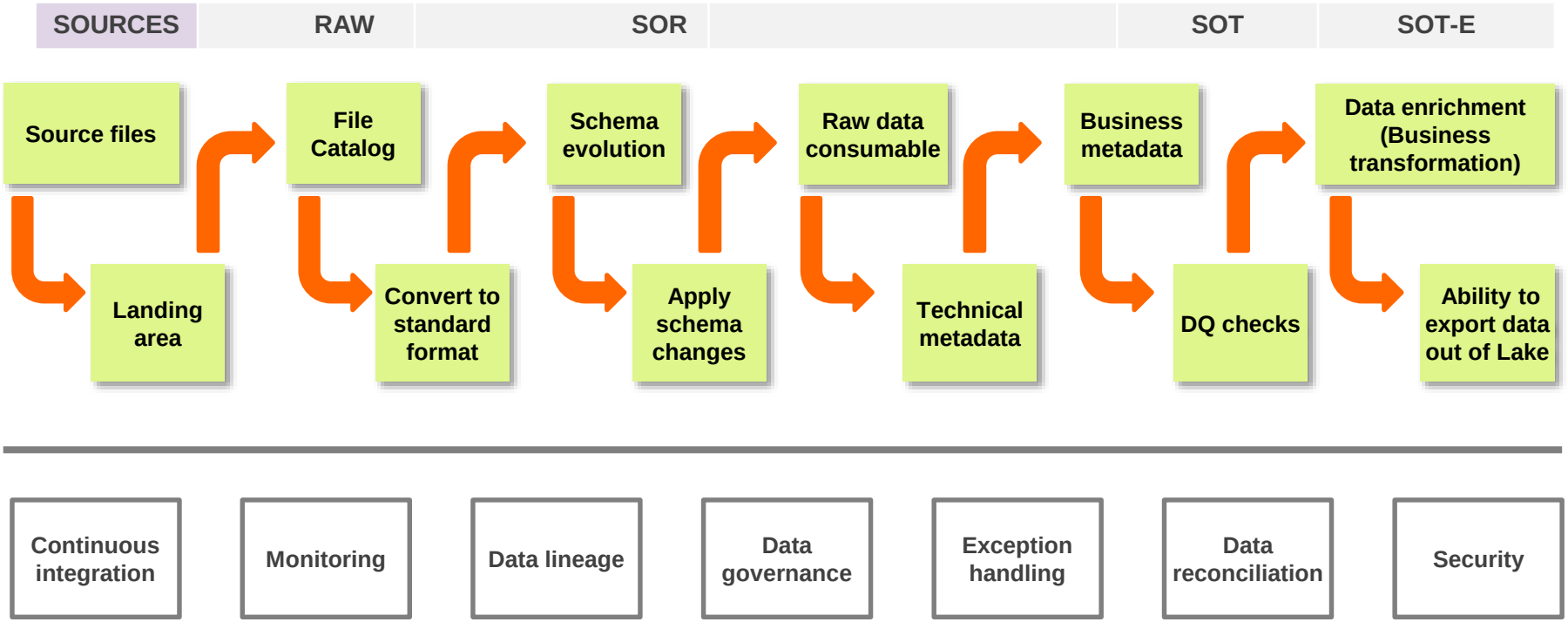
We are also extending the capability of DATAFLO CONTINUOUS DATA PIPELINE into the cloud



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
```

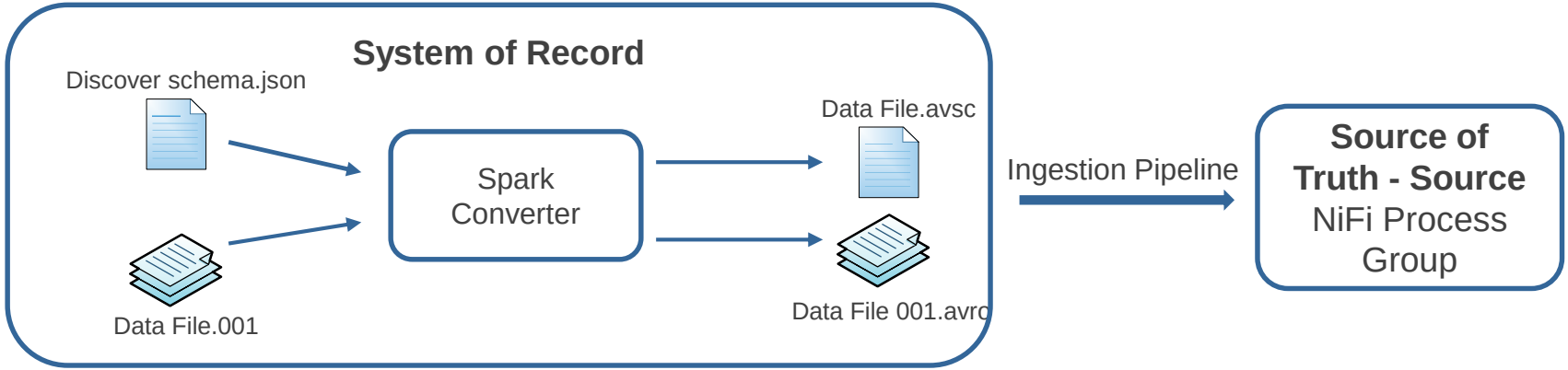
Example: Records with an array of Struct data types

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

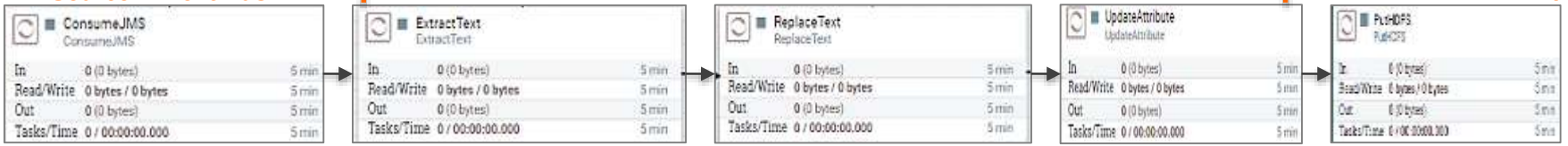
Destination: Hadoop

NiFi Ingestion Flow Version II

Complex logic, limited scale

Source : Event Bus

Destination: Hadoop



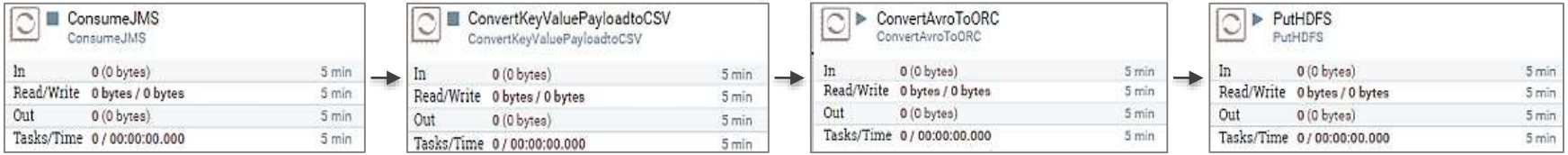
NiFi Ingestion Flow Version III

Seconds

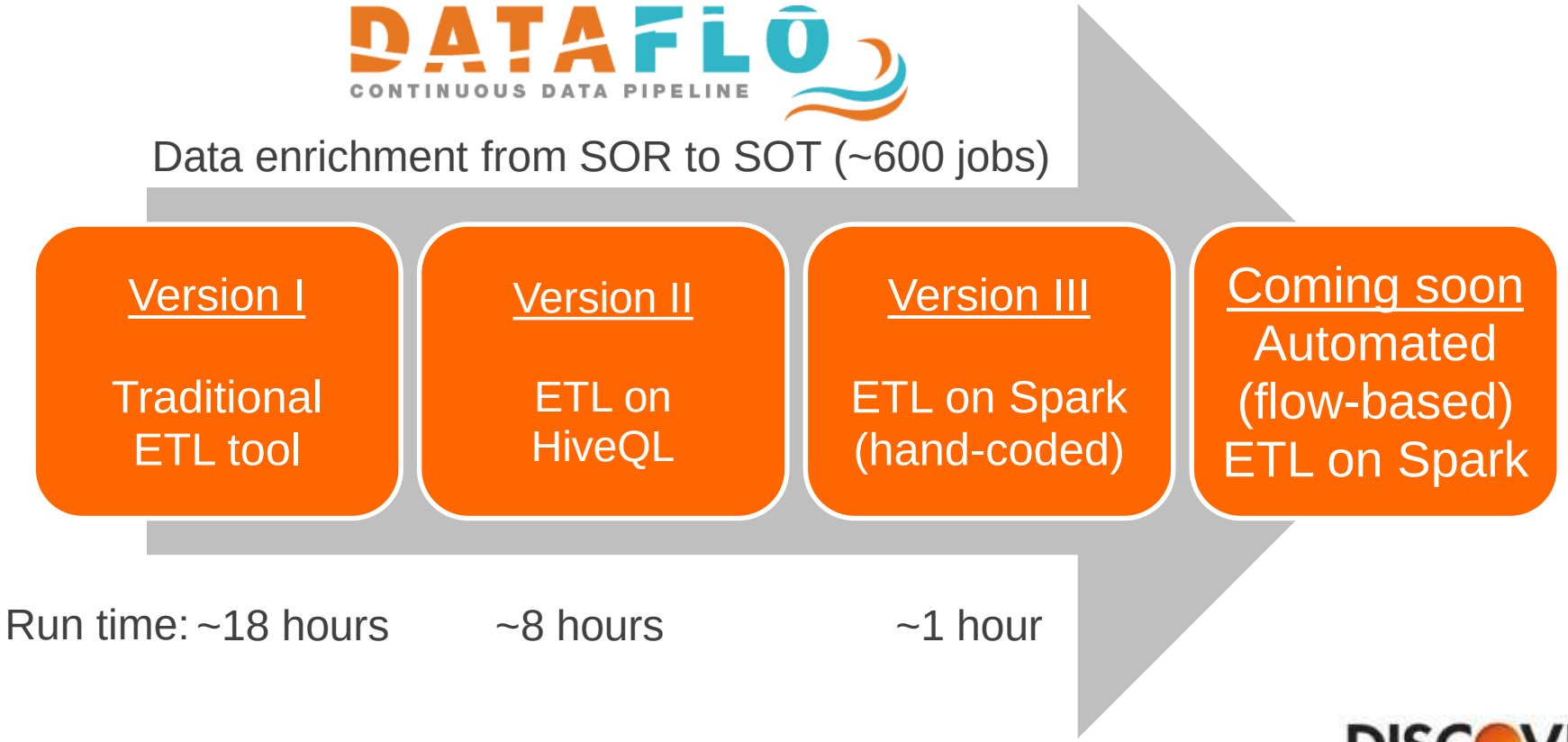
Custom NiFi processor developed in-house, reusable and scalable

Source : Event Bus

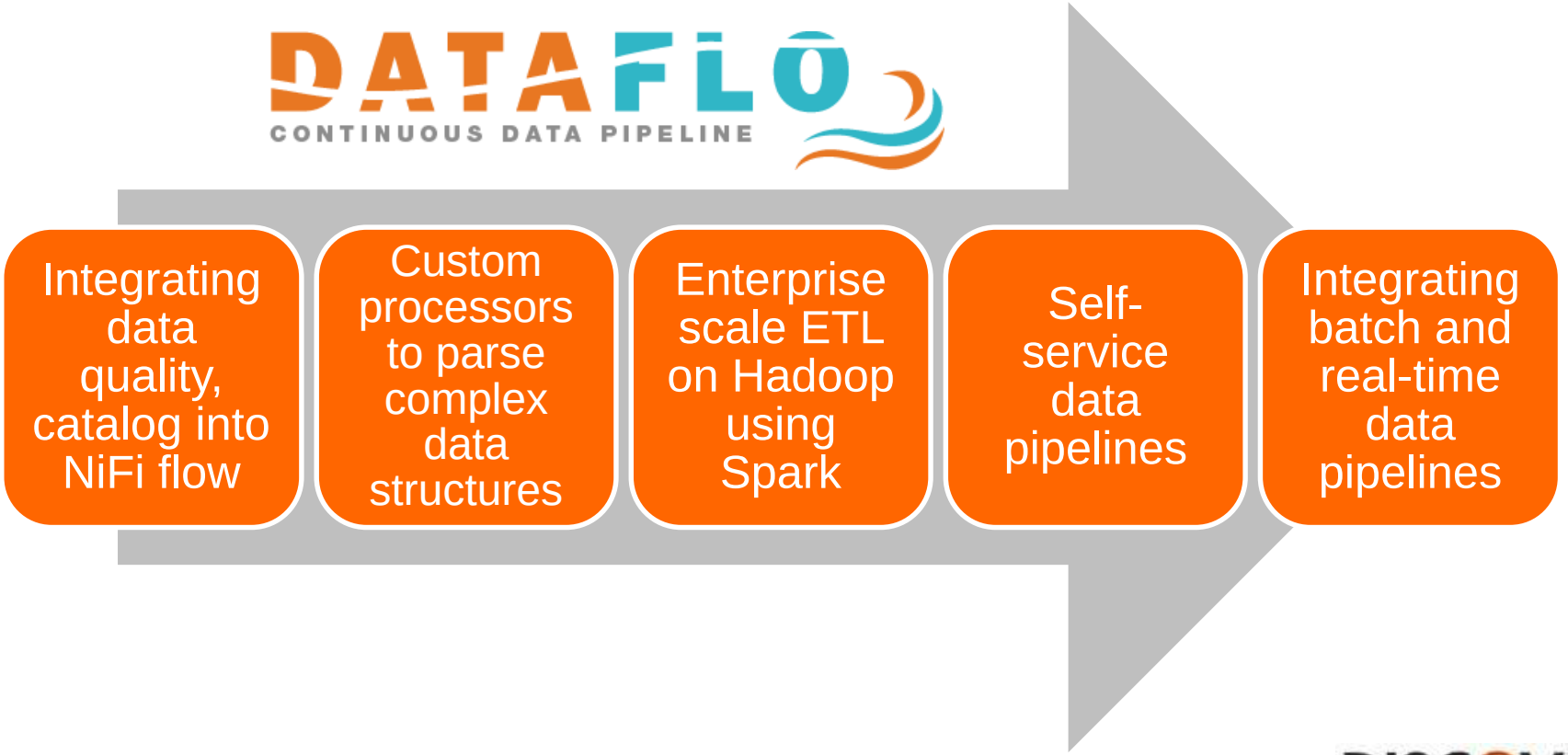
Destination: Hadoop



ETL on Hadoop progression



Upcoming enhancements to our data pipeline



DISCOVER®

Hiring Data Engineers

Q & A