When Big Data Ingestion Gets Complicated

Case studies in Global and Multi-Enterprise Ingestion

John Thielens
Chief Technology Officer, Cleo
NAVIGATION | CLIMATE

SOCIAL | MEDIA

BUSINESS | COMMERCE

MOBILE | COMMERCE

Files
Messages
Events
Rows
Complication – Push and Pull
A health care analytics company purchases machine data from a medical device firm to incorporate into their analysis. These large data sets need to be ingested into a Hadoop cluster and refreshed periodically. The selected technology for transfer between firms is FTP.

Who is going to be the FTP client and who will be the FTP server? How easy is it to change our minds? Either way, can we move from FTP to Hadoop without the latency and storage required to land a temporary copy?
Complication – Remote Copy
Remote Copy

Situation

- A marketing analytics firm ingests large data sets from their clients on a quarterly basis into Amazon S3 for analysis using Apache Spark.
- Some clients also use Amazon S3 to construct the data set, while others do not.

Complication

- Time (and money) can be saved using Amazon S3 native copy APIs instead of a naïve GET/PUT copy approach using an MFT engine.
- Can a uniform approach be used to manage the partner profile for each client, yet still allow the S3 optimization to be exploited?
Complication – Great Firewall
A CPG company collects weekly sales data from their distribution network in China.

The data sets are stored in Amazon S3 in China.

The company would like to aggregate this data with other data sets in the USA for analysis.

Extraction of the data sets from S3 in China is not as performant as expected.

Additionally, the TCP connections from China to the US are throttled, delaying the transmission.
Connect and Integrate

- B2B Integration
- Ground-to-Cloud
- Application Integration
- Big Data Integration and Ingestion

7,000+ Deployments
5x Growth in 5 Years
$275B Transactions / Day
900 Pre-Built Connectors
3+ Terabytes / Hour
99% Fortune 10 Renewals

Acceleration
- Streaming
- Linear Scaling

Big Data
- AS2
- FTP
- Network Fax
- EDI
- cloudera
- SAP
- Amazon Web Services
- Microsoft Azure
- Oracle
- JD Edwards
- Banking
- Retail
- Healthcare
- Manufacturing
- E-commerce
- Telco
- Energy
- Government
- Aerospace
- Transportation

AS2, FTP, and 18 more

Connect and Integrate