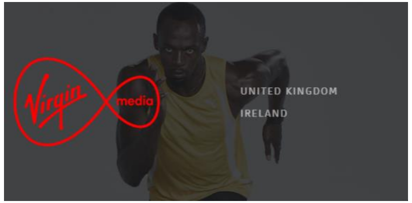
DATA ACQUISITION AUTOMATION FOR NIFI
IN A HYBRID CLOUD ENVIRONMENT

The path towards DataOps

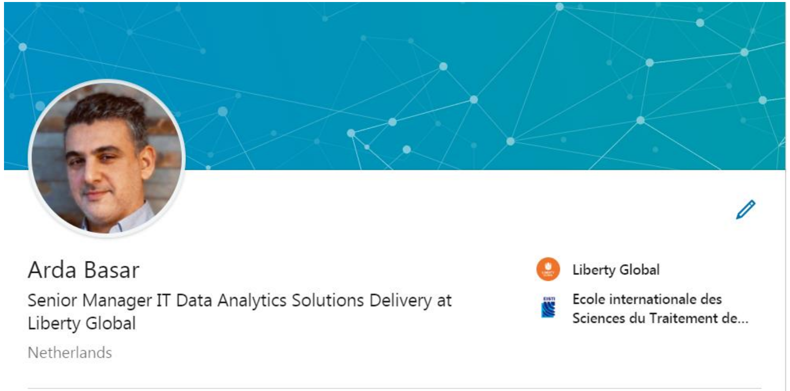
LIBERTY GLOBAL

- Liberty Global is the world’s largest international TV and broadband company, with operations in 10 European countries.*
- 21 million customers subscribing to 45 million TV, broadband internet and telephony services, 6 million mobile subscribers.*
- WiFi service through 12 million access points across our footprint.*
- Liberty Global owns 50% of VodafoneZiggo, a joint venture in the Netherlands with 4 million customers subscribing to 10 million fixed-line and 5 million mobile services
- Significant investments in ITV, All3Media, ITI Neovision, LionsGate, the Formula E racing series and several regional sports networks.

* The figures included in this paragraph include both the continuing and discontinued operations that we owned on December 31, 2018



SPEAKERS



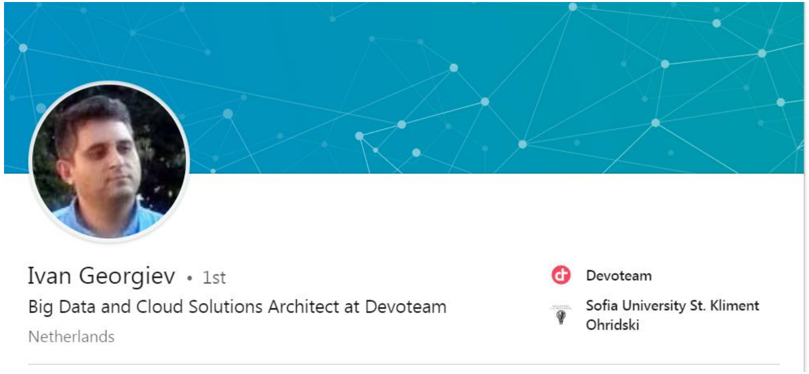
Arda Basar

Senior Manager IT Data Analytics Solutions Delivery at Liberty Global

Netherlands

Liberty Global

Ecole internationale des Sciences du Traitement de...



Ivan Georgiev • 1st

Big Data and Cloud Solutions Architect at Devoteam

Netherlands

Devoteam

Sofia University St. Kliment Ohridski

linkedin.com/in/arbasar

linkedin.com/in/bigbaobab

“WE NEED TO MAKE IT EASIER FOR THE PROJECTS TO COME ON-BOARD. MY WORRY IS THAT IF WE PRESENT THE PROJECTS WITH SUCH LENGTHY PROCESS, THIS MIGHT PREVENT THEM FROM USING THE PLATFORM. ”

HEAD OF ADVANCED ANALYTICS

“GDPR FORCES US TO TAKE SIGNIFICANTLY MORE CONSCIOUS DECISIONS ON WHAT DATA WE COLLECT, HOW WE USE THE DATA, HOW LONG WE KEEP THE DATA AND WILL LEAD TO IMPROVE DATA QUALITY”

DATA PROTECTION OFFICER

“ […] I WOULD EXPECT […] THE DATA WITHIN THE PLATFORM IS PROTECTED BY DEFAULT, WHETHER THAT MAY BE (PSEUDO)ANONYMIZATION, TOKENISATION OR ENCRYPTION OF FIELDS OR OTHER MECHANISMS. ”

GLOBAL SECURITY ARCHITECT

DEFINITION

Data Acquisition

from New Horizons for a Data-Driven Economy: A Roadmap for Usage and Exploitation of Big Data in Europe

[…] Data acquisition has been understood as the process of **gathering**, **filtering**, and **cleaning data** before the data is put in a data warehouse or any other storage solution.

DEFINITION

DataOps

From Wikipedia, the free encyclopedia

DataOps is an **automated**, **process-oriented** methodology, used by analytic and data teams, to **improve the quality** and **reduce the cycle time** of data analytics.

[…] From a process and methodology perspective, DataOps applies **Agile software development**, **DevOps** and the statistical process control used in **lean manufacturing**, to data analytics.

PARTNERS

Key Partners

cloudera

HORTONWORKS®

Microsoft

ATLASSIAN

SecuPi

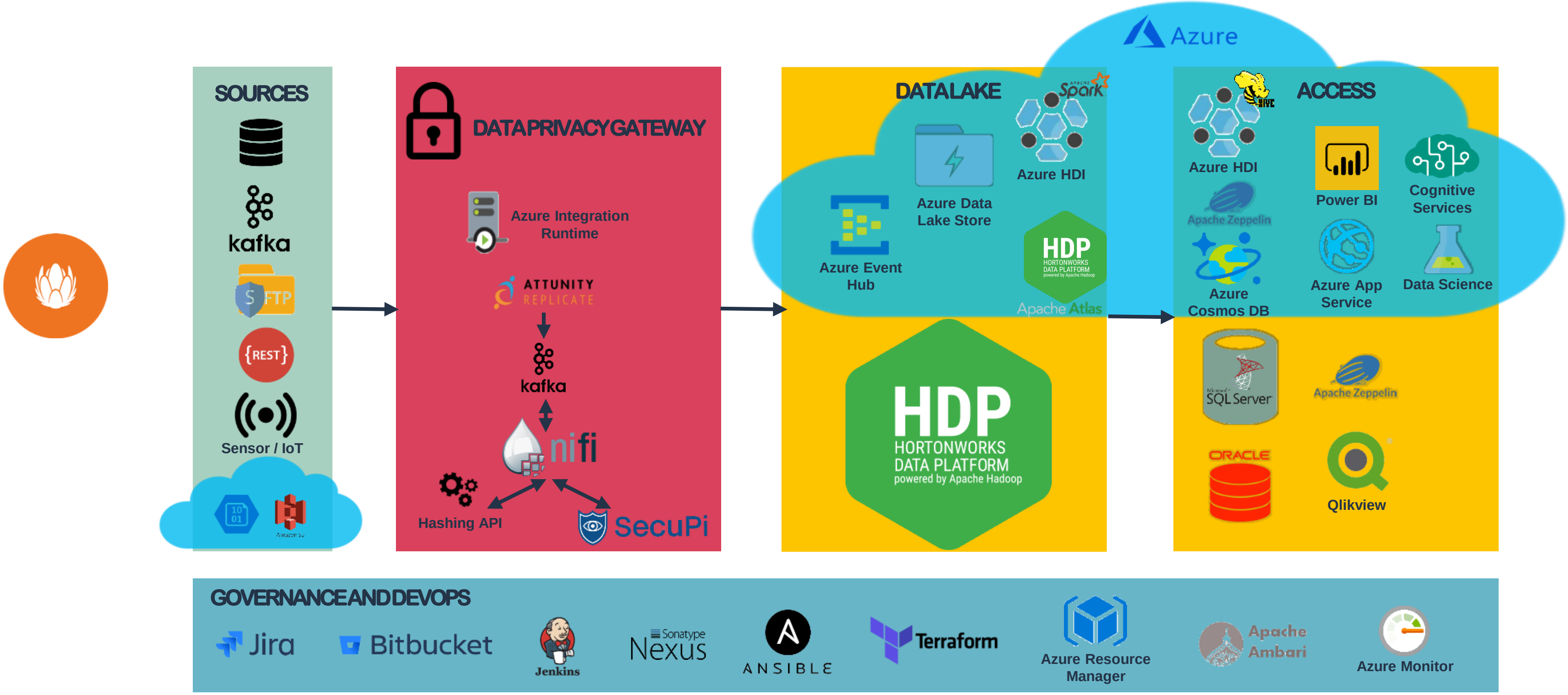
Key Integration Partners

devoteam

Platform, Engineering and Support

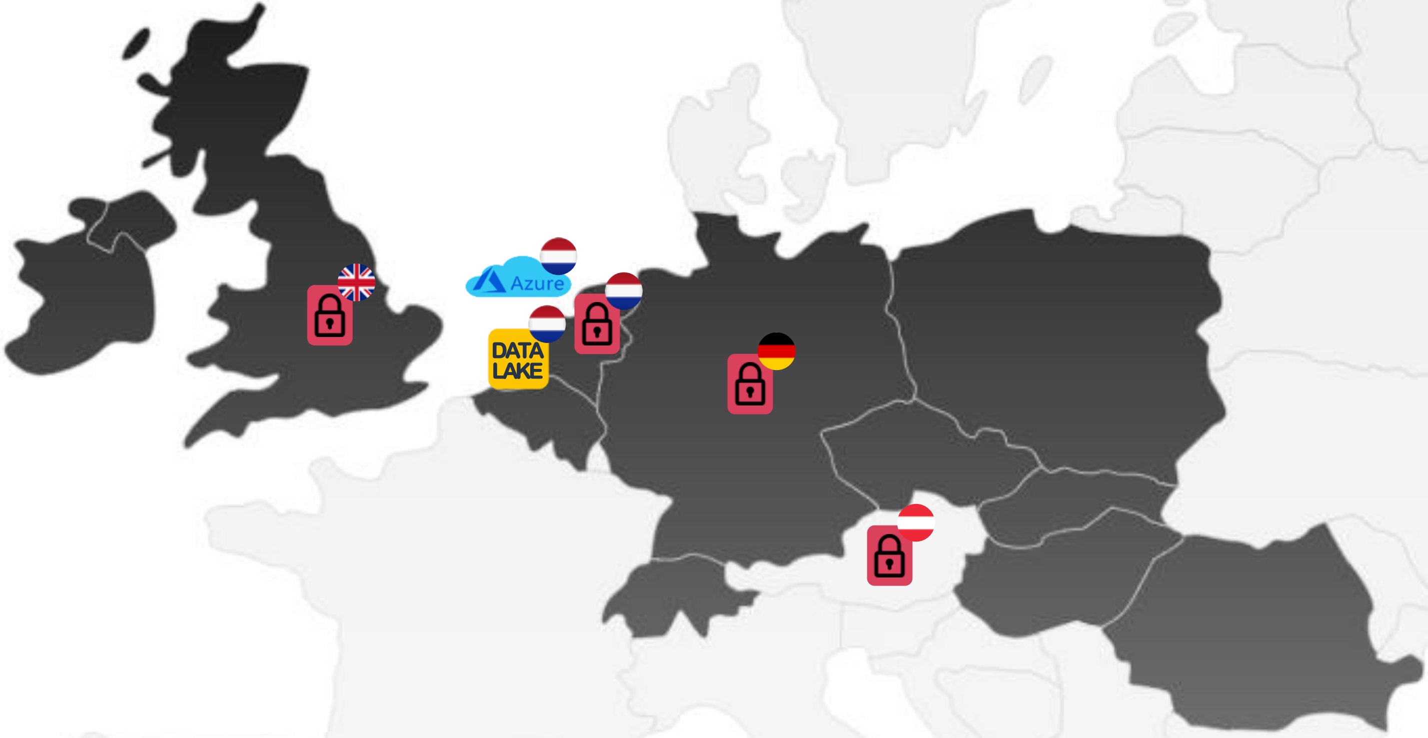
THE HYBRID CLOUD DATA PLATFORM

Reference Architecture



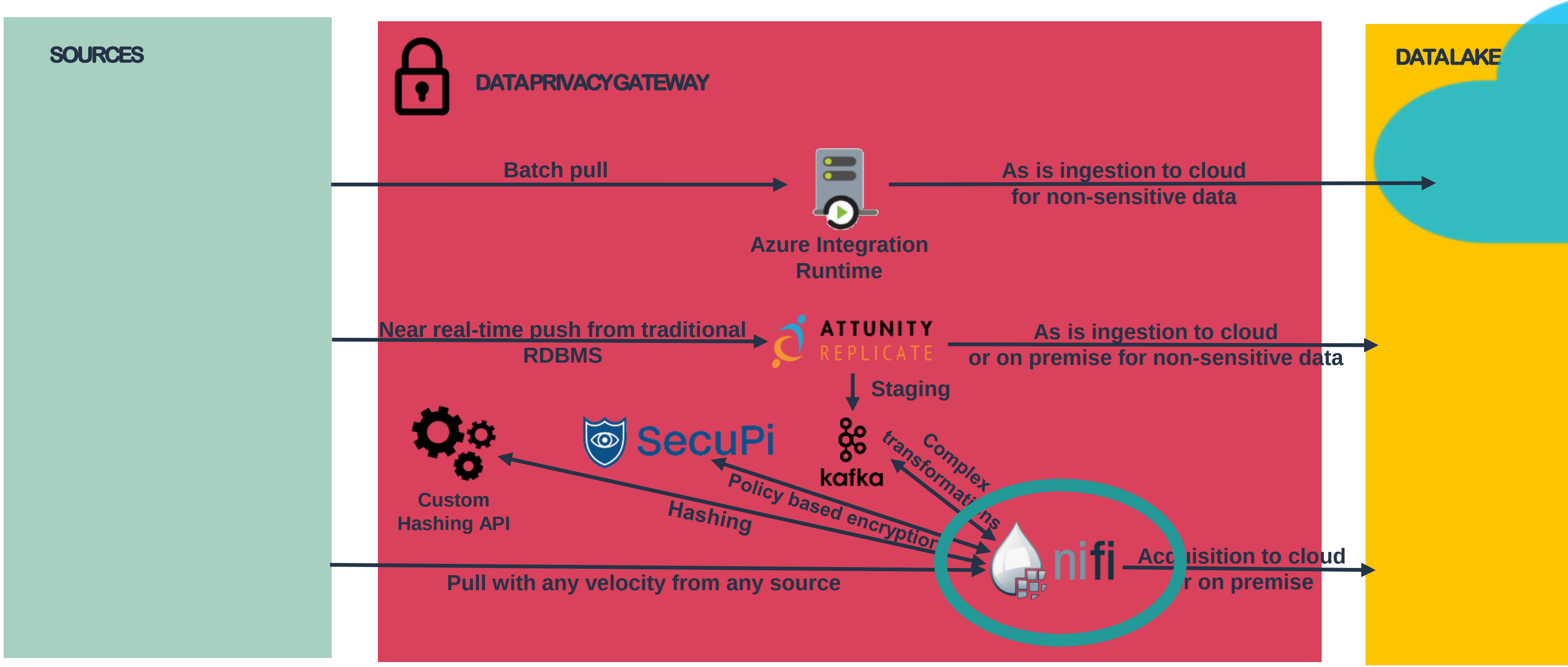
THE HYBRID CLOUD DATA PLATFORM

Infrastructure location



THE HYBRID CLOUD DATA PLATFORM

Data ingestion and acquisition though the Data Privacy Gateway


SOURCES
DATAPRIVACYGATEWAY
DATALAKE
Batch pull
Azure Integration Runtime
As is ingestion to cloud for non-sensitive data
Near real-time push from traditional RDBMS
ATTUNITY REPLICATE
As is ingestion to cloud or on premise for non-sensitive data
Staging
Custom Hashing API
SecuPi
kafka
Complex transformations
Policy based encryption
Hashing
nifi
Acquisition to cloud or on premise
Pull with any velocity from any source

THE HYBRID CLOUD DATA PLATFORM

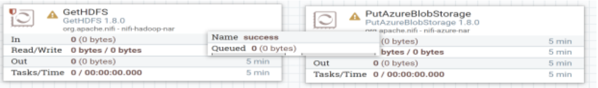
Data ingestion and acquisition though the Data Privacy Gateway

- The Data Privacy Gateway is used to acquire and ingest data into the Data Lake.
- It can deliver data to on-premises and to the cloud Data Lake.
- The gateway is responsible not only for transport, but also performs sensitive data management by applying anonymization or pseudonymization techniques on sensitive data. It removes personal data and ensures the Data Lake is fully GDPR compliant.

DATA ACQUISITION FLOW EVOLUTION

WITH nifi

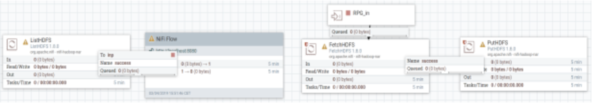
June 2018


GetHDFS
GetHDFS 1.8.0
org.apache.nifi - nifi-hadoop-nar
In 0 (0 bytes)
Read/Write 0 bytes / 0 bytes
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min
Name success
Queued 0 (0 bytes)
PutAzureBlobStorage
PutAzureBlobStorage 1.8.0
org.apache.nifi - nifi-azure-nar
(0 bytes) 5 min
bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min

NiFi Flow foundation

- Created about 200 streams using Get-Put pattern and provide input to Data Privacy team
- Acquisition, management and save capacity is limited. How to scale?

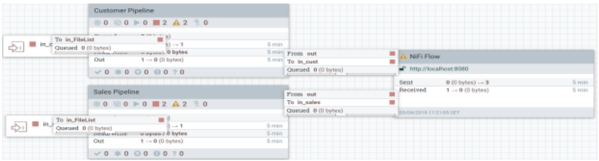
August 2018



Scalability

- Redesigned, regression tested all 200 streams to List-Distribute-Fetch pattern with Remote Process Group.
- Each RPG and associated input ports allocate resources – processes. How to limit resources? RPG requires root level input port – clutters the design.

September 2018

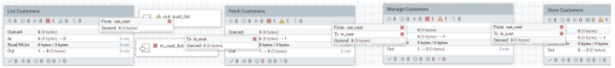


Resource usage optimization

- Redesigned, regression tested and all 200 streams to share single RPG.
- Privacy management framework not yet selected. How to de-identified data?

October 2018

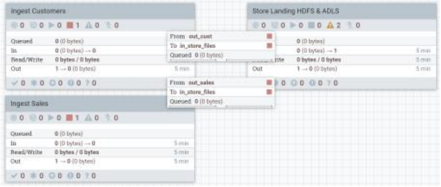
Templates and GDPR



- For each “Acquire”, “Manage” and “Store” step, a NiFi template was created. For the de-identification, specific NiFi processors were implemented, all 200 streams were redesigned, regression tested and deployed to share single RPG.
- On-premise hardware is delayed. How to change the target to Azure Blob storage instead of HDFS?

November 2018

Multiple targets



- Extracted store step from each pipeline into a shared step, capable of writing to two targets. Updated, regression tested and deployed all 200 streams to share the “Store” step.
- A major NiFi version (1.8) is released, fixing the query (Oracle) database issue. In addition, it makes RPG redundant for load balancing. Also, template instances are altered manually by developers.

December 2018

Version Control

- All steps from all 200 streams placed under NiFi Registry version control. Updated, regression tested and deployed.
- NiFi registry doesn’t support branching.
- Security restriction: cannot connect NiFi registry to DEV and PROD at the same time, no human access to NiFi UI on PROD.

February 2019

CI/CD

APACHE NIFI registry

nifi

python™

Bitbucket

- Implemented NiFi Continuous Integration/Continuous Delivery pipeline, using NiFi REST API, orchestrated by Jenkins. Using Bitbucket as additional version control system.

- New Requirement: Change source systems for all streams.
- New Requirement: 3rd party privacy framework (SecuPi).
- New Requirement: Prospective customers require rapid data streams engineering for exploration.
- Challenge: NiFi templates and NiFi registry have limitations on variables and processor properties.
- Challenge: Dataworks Summit in Barcelona is approaching!

Automation

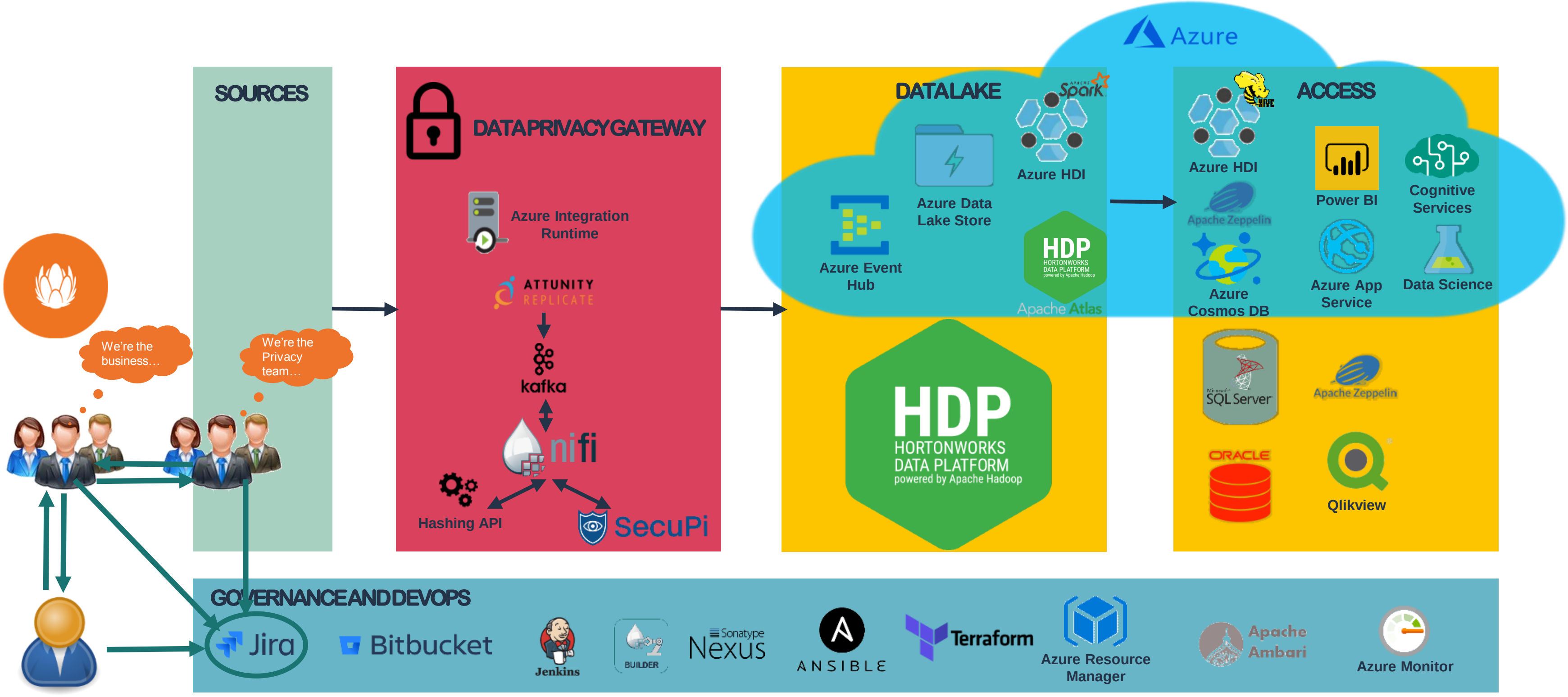
Introducing # NiFi Builder

- Command-line interface for declarative NiFi streams construction.
- Declarations use JSON as descriptor format.
- Support for multiple target platforms – e.g. NiFi, Azure Data Factory etc.
- Rapid stream construction, easy maintenance.
- Eliminate waste by reducing repetitive manual work.

NIFI BUILDER DEMO

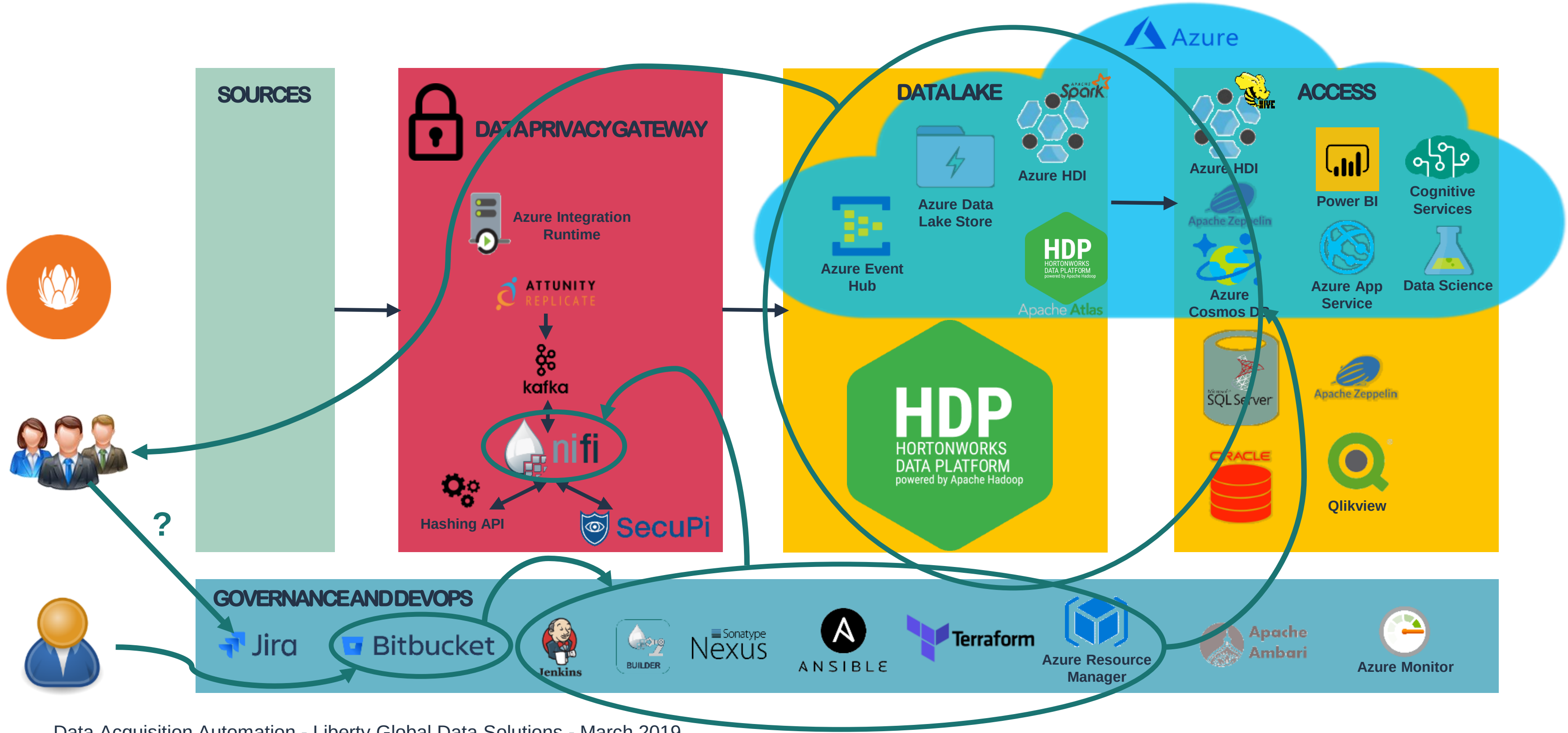
BEFORE DATAOPS

The business requirement gathering process



BEFORE DATAOPS

The development / deployment process



THE PATH TOWARDS DATAOPS

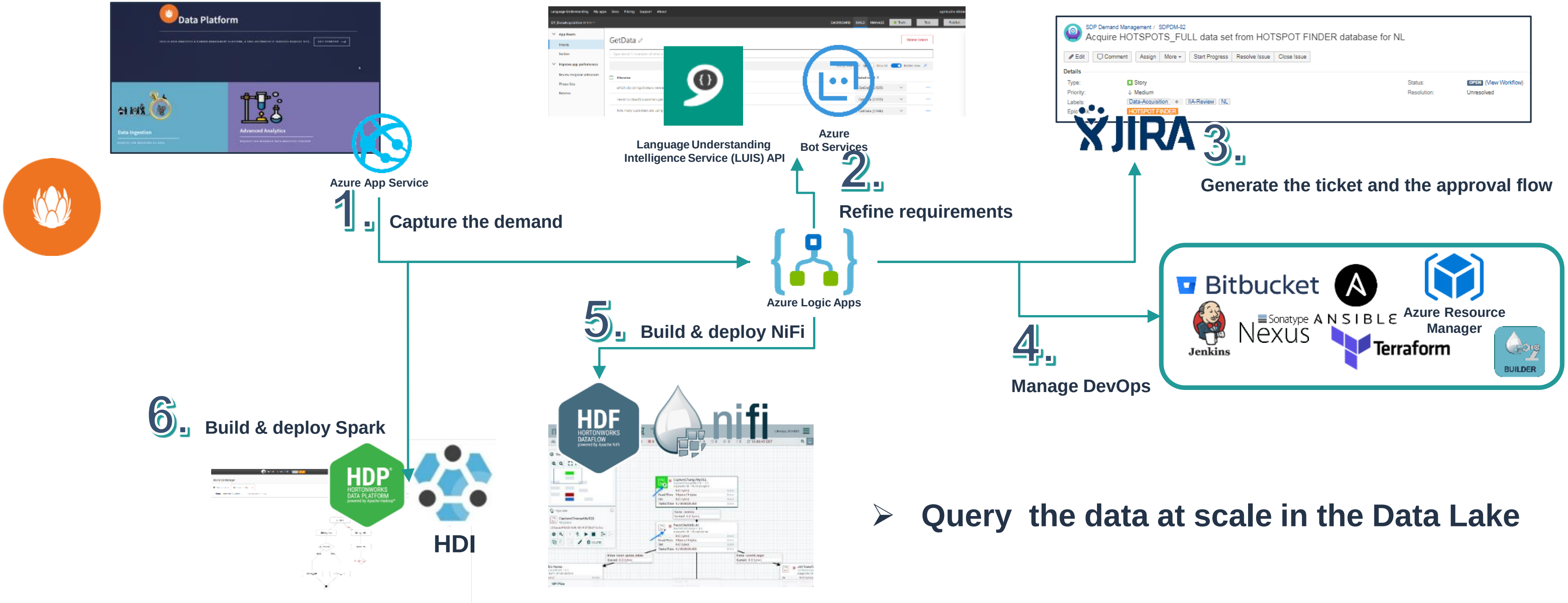
Challenges

- It's difficult to capture requirements in a consistent and standardized manner as projects are driven by multiple business units and affiliates.
- Approvals for budget, data privacy, legal, IT teams, Network and operations are often done in isolated environments such as e-mail, SharePoint, JIRA, Excel and others.
- Filling-in forms for all the requirements and approvals is a repetitive and lengthy task.
- Most business users don't know in which technical system to find the data they need.

Solution: All the Governance and DevOps tools should be integrated to enforce the business processes

DATAOPS PLATFORM – DATA ACQUISITON

Solution Process Flow



SDP Demand Management / SDPDM-82

Acquire HOTSPOTS_FULL data set from HOTSPOT FINDER database for NL

Edit | Comment | Assign | More | Start Progress | Resolve Issue | Close Issue

Details

Type: Story

Priority: Medium

Labels: Data-Acquisition IIA-Review NL

Epic Link: HOTSPOT FINDER

Status: OPEN (View Workflow)

Resolution: Unresolved

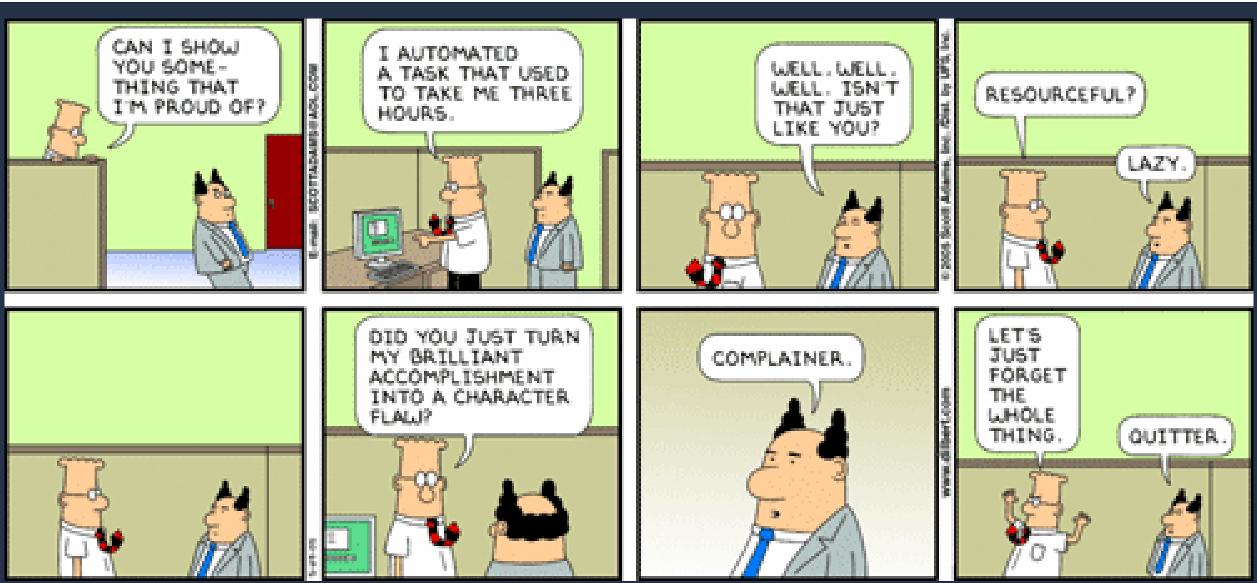
➢ Query the data at scale in the Data Lake

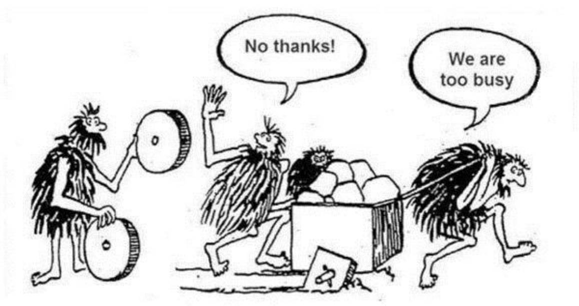
DATA ACQUISITION DATAOPS DEMO

WHAT'S NEXT?

Ideas for becoming a DataOps Factory

- Integration with the Business Glossary, Application Landscape Management.
- Automate connectivity checks and requests.
- Monitoring and support.
- DataOps Platform closed beta.
- Review business process based on experience.
- R&D for complex service automation: Data Integration, Data Quality, Dashboards.


CAN I SHOW YOU SOME-THING THAT I'M PROUD OF?
I AUTOMATED A TASK THAT USED TO TAKE ME THREE HOURS.
WELL, WELL, WELL. ISN'T THAT JUST LIKE YOU?
RESOURCEFUL?
LAZY.
DID YOU JUST TURN MY BRILLIANT ACCOMPLISHMENT INTO A CHARACTER FLAW?
COMPLAINER.
LET'S JUST FORGET THE WHOLE THING.
QUITTER.
www.dilbert.com


No thanks!
We are too busy

Q & A