BigData
UN Global Working Group

4TH INTERNATIONAL CONFERENCE ON

BIG DATA

for Official Statistics

8-10 NOVEMBER 2017
BOGOTA, COLOMBIA

Trade & Transport Data Lake
(a proof of concept)

GOBIERNO DE COLOMBIA MINTIC DANE INFORMACIÓN ESTRATÉGICA

Vision for Trade Data Lake

Data from new sources such as UPU (postal), ICAO (aviation), vessel tracking, etc, possibly in streams, to be included along with traditional trade statistics plus the underlying administrative data sources (such as customs declarations or shipping manifests) for better in-depth, timely and accurate analysis/forecast of international trade.

The comparative advantage over existing data would be timelier and richer analysis with possibilities of nowcasting and accurate short term forecasting

Problem statement

The introduction of new heterogeneous streams of data from various new sources demands

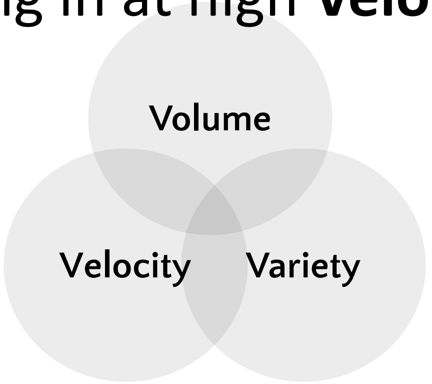
- elastic storage capacity that is flexible and more reliable
- real-time analytical capabilities (schema-on-read)
- convenience to do visualizations across these different data sources (structured/unstructured, processed/unprocessed)
- access to data anywhere/anytime through secure channel

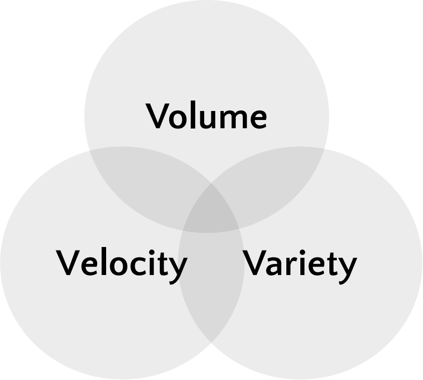
Addressing the challenge

- Reliable elastic storage requirements can be met by moving to cloud-based solutions
- Real-time analytical capability can be realized with interactive query tools (big data analytics)
- Visualization across divergent datasets can possibly achieved through SaaS (Software as a service) tools

Features of big data analytics

- Possible to store and analyze huge **volumes** of structured/unstructured data that has **Variety** in nature and flowing in at high **Velocity**.





Volume

Velocity

Variety

- When the existing platforms/infrastructures cannot withstand the three Vs, we move to Cloud based platforms such as AWS.
- When the existing analytical solutions cannot withstand the three Vs, we move to Big data analytics such as Hadoop.

Types of cloud architectures

IaaS/PaaS

(Infrastructure as a service/Platform as a service)

The cloud vendor provides the service up to OS level (IaaS) or application level (PaaS) and we manage the rest.

- The required service is not available in the cloud (such as custom-applications built in house).
- bound to specific technologies/tools

Serverless (SaaS)

(Software as a service)

No servers, no maintenance and no licensing overheads. Desired functionalities are in the form of microservices (or nanoservices). No provisioning of underlying platforms such as OS, application, etc.

- The required service is readily available on the cloud such as data visualizations, api gateway, etc.
- Not bound to any tool/technology.

Major cloud players

Microsoft Azure

amazon web services

Google Cloud Platform

Major hadoop players

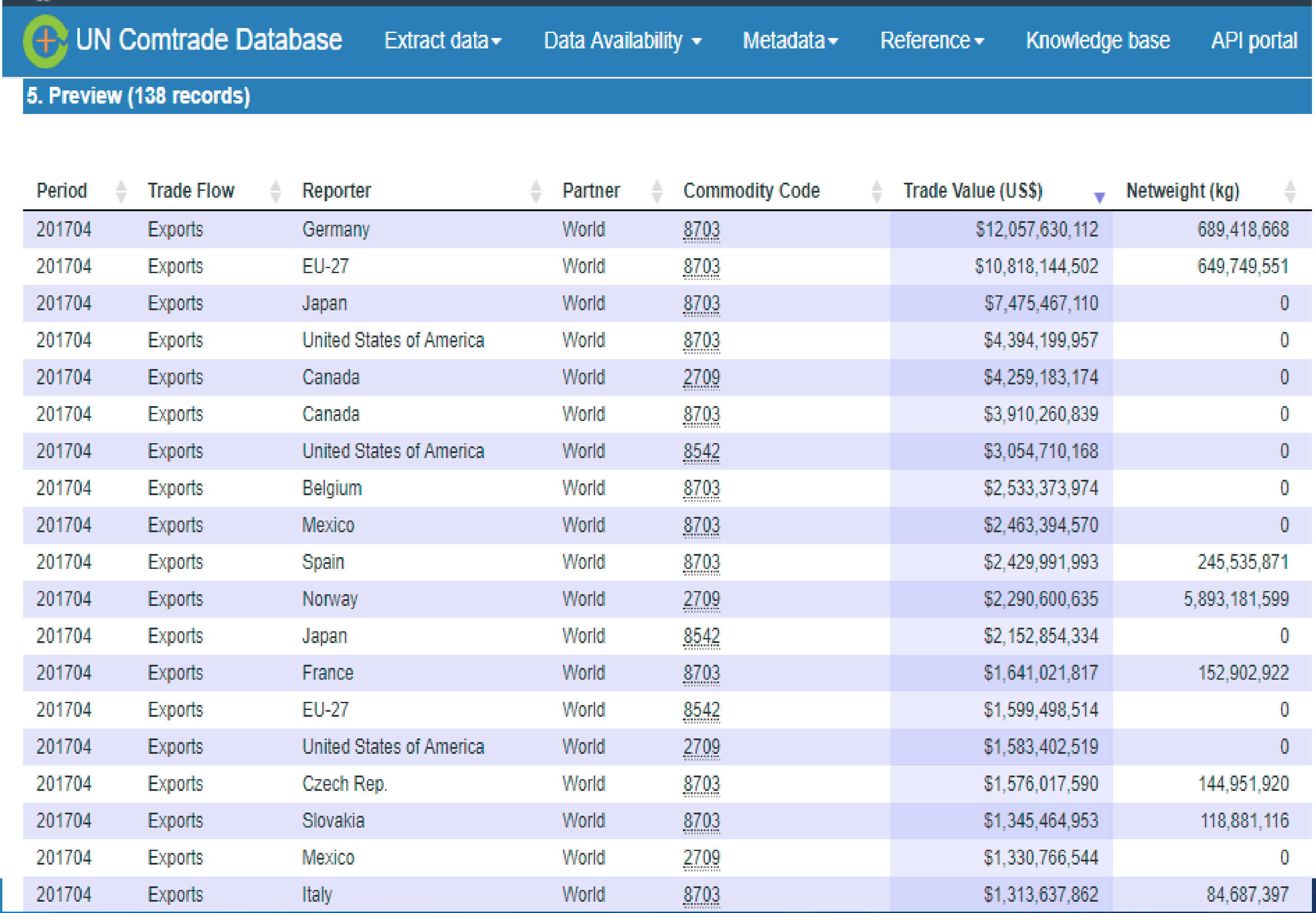
HORTONWORKS®
POWERING THE FUTURE OF DATA™

MAPR®

cloudera

Inside the Trade Data Lake

Trade Data (National)



5. Preview (138 records)

Period	Trade Flow	Reporter	Partner	Commodity Code	Trade Value (US$)	Netweight (kg)
201704	Exports	Germany	World	8703	$12,057,630,112	689,418,668
201704	Exports	EU-27	World	8703	$10,818,144,502	649,749,551
201704	Exports	Japan	World	8703	$7,475,467,110	0
201704	Exports	United States of America	World	8703	$4,394,199,957	0
201704	Exports	Canada	World	2709	$4,259,183,174	0
201704	Exports	Canada	World	8703	$3,910,260,839	0
201704	Exports	United States of America	World	8542	$3,054,710,168	0
201704	Exports	Belgium	World	8703	$2,533,373,974	0
201704	Exports	Mexico	World	8703	$2,463,394,570	0
201704	Exports	Spain	World	8703	$2,429,991,993	245,535,871
201704	Exports	Norway	World	2709	$2,290,600,635	5,893,181,599
201704	Exports	Japan	World	8542	$2,152,854,334	0
201704	Exports	France	World	8703	$1,641,021,817	152,902,922
201704	Exports	EU-27	World	8542	$1,599,498,514	0
201704	Exports	United States of America	World	2709	$1,583,402,519	0
201704	Exports	Czech Rep.	World	8703	$1,576,017,590	144,951,920
201704	Exports	Slovakia	World	8703	$1,345,464,953	118,881,116
201704	Exports	Mexico	World	2709	$1,330,766,544	0
201704	Exports	Italy	World	8703	$1,313,637,862	84,687,397

Courtesy: upu.int

Postal Data (National)

CT_Address.CT_Location.code	String
CT_Address.CT_Location.name	String
CT_m307_supported_messages.T_Item.attributes_version_number	String
CT_m307_supported_messages.T_Item.customs_procedure	String
CT_m307_supported_messages.T_Item.dangerous_good_hazard_class	String
CT_m307_supported_messages.T_Item.dangerous_good_type_code	String
CT_m307_supported_messages.T_Item.declared_gross_weight	String
CT_m307_supported_messages.T_Item.measured_gross_weight	String
CT_m307_supported_messages.T_Item.nature_of_transaction_code	String

Courtesy: upu.int

BigData
UN Global Working Group

AIS Vessel Tracking Data

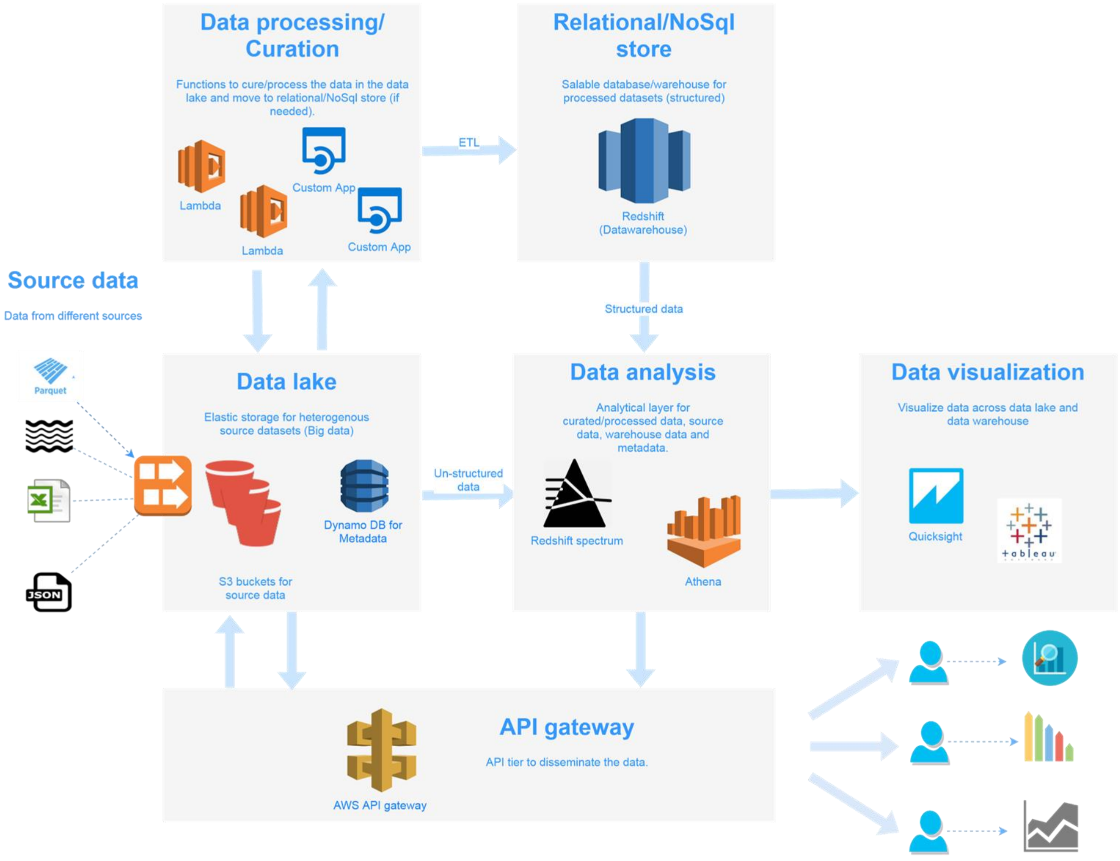
FIELD	DESCRIPTION
TIMESTAMP	Time of ship position detection / reception (in UTC)
MMSI	Ship's MMSI number sent with the AIS notification
Lat	Latitude of the ship position (in decimal degrees)
Lon	Longitude of the ship position (in decimal degrees)
"Speed over ground"	Speed over ground (in knots)
"Course over ground"	Course over ground (in degrees)
Heading	True heading (in degrees (0-359))
"IMO number"	Ship's IMO number sent with the AIS notification
Shipname	Ship's vessel name sent with the AIS notification
Callsign	Ship's Callsign sent with the AIS notification
"Type of ship"	Ship type
Draught	Maximum Present Static Draught (in meters)
Destination	Destination

Courtesy: Statistics Netherlands (from ESS bigdata for official statistics)

In addition:

- Freight Aviation Data (ICAO)
- Shipping Manifests
- Advanced Export Declarations
- ASYCUDA records

Data Architecture based on AWS Cloud solutions



Data processing/ Curation

Functions to cure/process the data in the data lake and move to relational/NoSql store (if needed).

Lambda

Custom App

Lambda

Custom App

ETL

Relational/NoSql store

Salable database/warehouse for processed datasets (structured)

Redshift (Datawarehouse)

Structured data

Source data

Data from different sources

Parquet

JSON

Data lake

Elastic storage for heterogenous source datasets (Big data)

S3 buckets for source data

Dynamo DB for Metadata

Un-structured data

Data analysis

Analytical layer for curated/processed data, source data, warehouse data and metadata.

Redshift spectrum

Athena

Data visualization

Visualize data across data lake and data warehouse

Quicksight

tableau

API gateway

API tier to disseminate the data.

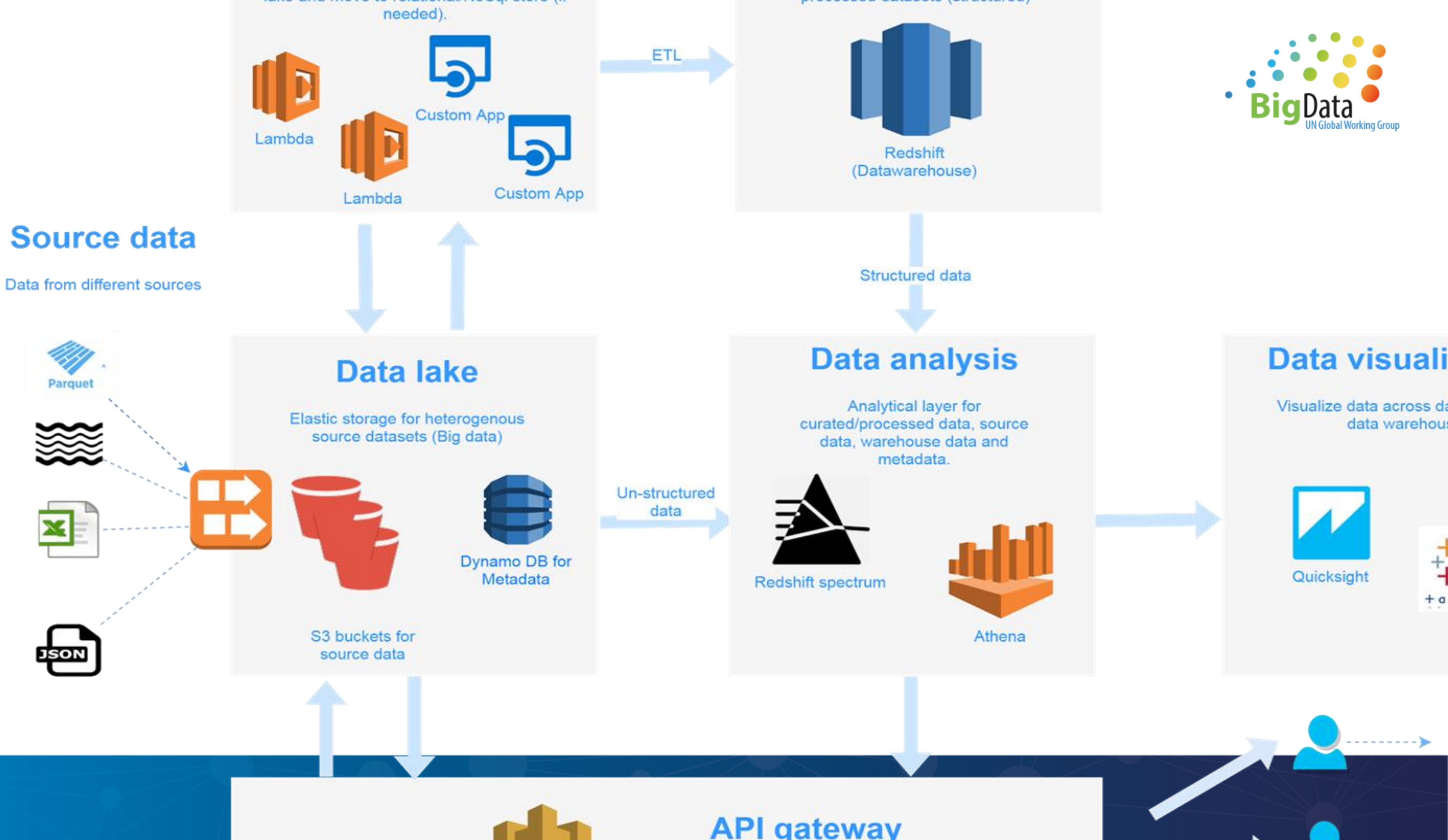
AWS API gateway

Data Lake is AWS's elastic storage layer to dump multi data sources

These data can be curated and stored at AWS data warehouse service (Redshift)

Or can be directly analyzed in analytical layer

Raw data at Data lake and Refined data from analytical layer can be distributed through API or be visualized



lake and move to relational/NoSQL store (if needed).

processed datasets (structured)

Source data

Data from different sources

Data lake

Elastic storage for heterogenous source datasets (Big data)

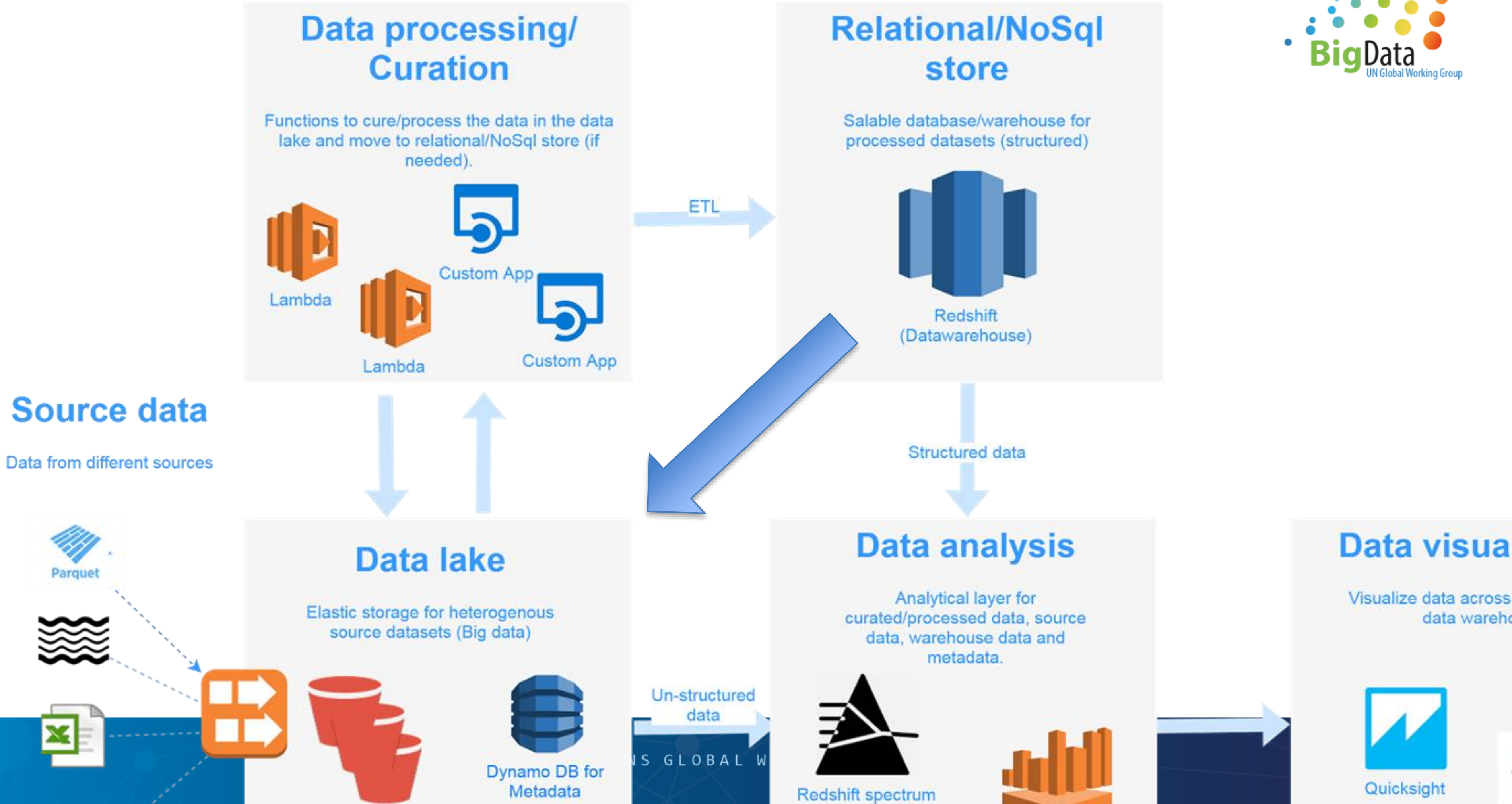
Data analysis

Analytical layer for curated/processed data, source data, warehouse data and metadata.

Data visuali

Visualize data across da
data warehous

API gateway



Data processing/ Curation

Functions to cure/process the data in the data lake and move to relational/NoSql store (if needed).

Relational/NoSql store

Salable database/warehouse for processed datasets (structured)

Source data

Data from different sources

Data lake

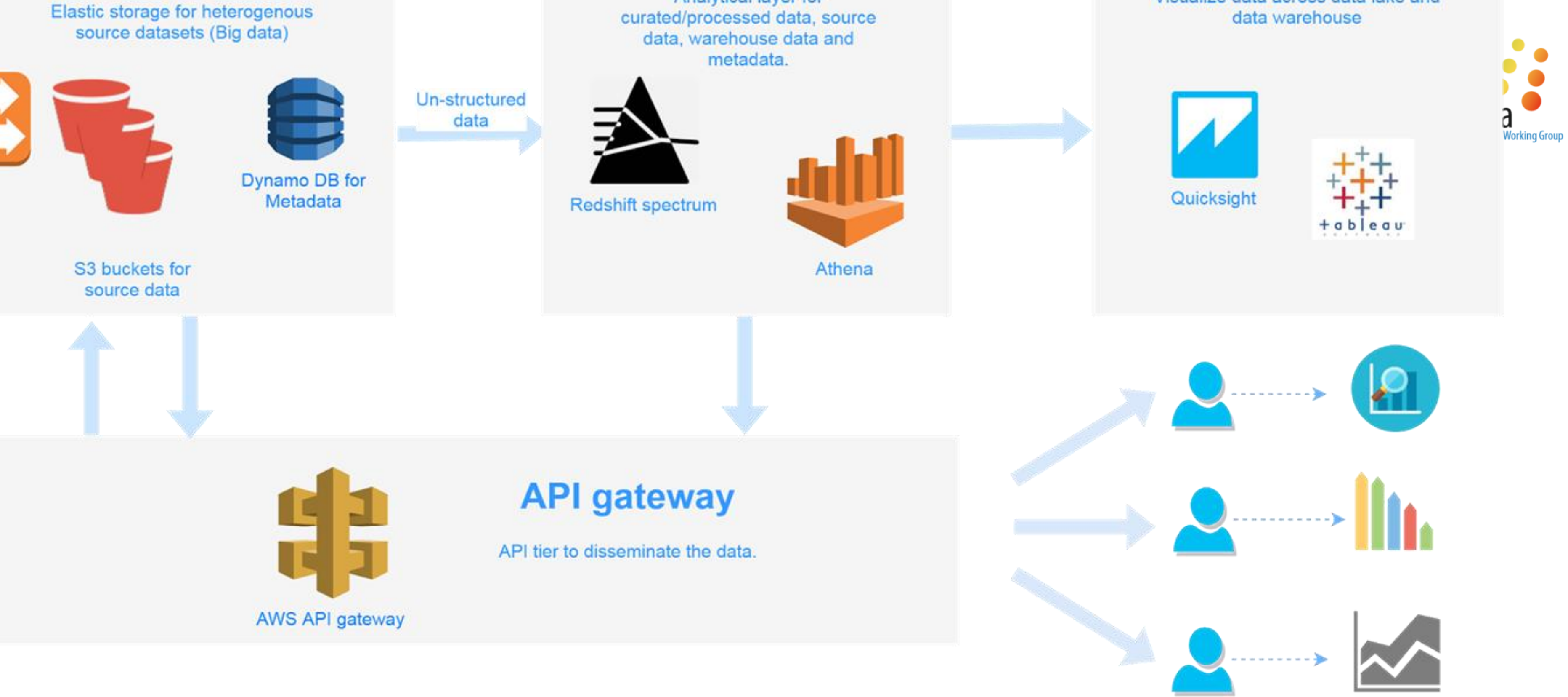
Elastic storage for heterogenous source datasets (Big data)

Data analysis

Analytical layer for curated/processed data, source data, warehouse data and metadata.

Data visua

Visualize data across data wareh



Elastic storage for heterogenous source datasets (Big data)

S3 buckets for source data

Dynamo DB for Metadata

Un-structured data

Analytical layer for curated/processed data, source data, warehouse data and metadata.

Redshift spectrum

Athena

Visualize data across data lake and data warehouse

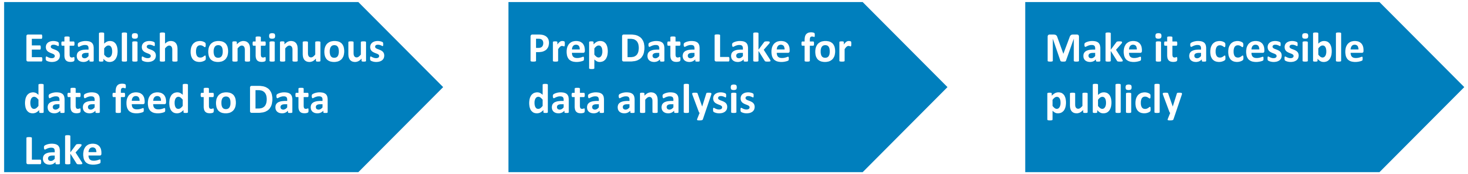
Quicksight

API gateway

API tier to disseminate the data.

AWS API gateway

Next steps…



Establish continuous data feed to Data Lake

Prep Data Lake for data analysis

Make it accessible publicly

THE UNITED NATIONS GLOBAL WORKING GROUP ON BIG DATA

Thanks!

Any questions ?

#UNBIGDATA2017

GOBIERNO DE COLOMBIA MINTIC DANE INFORMACIÓN ESTRATÉGICA