CERN

HANDLING 1PB OF IOT DATA TO CONTROL THE LARGEST SCIENTIFIC INSTRUMENT WITH ORACLE AUTONOMOUS DATA WAREHOUSE

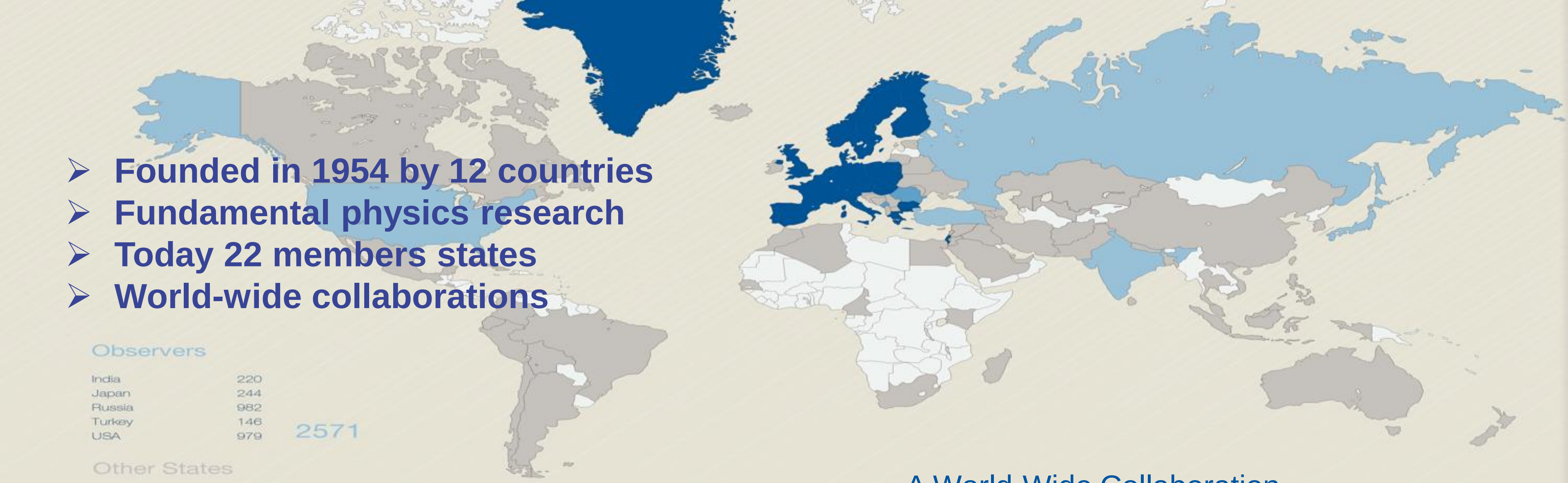
Manuel Martín Márquez, Senior Project Leader
Sebastien Masson, Oracle DBA
CERN – IT Database Services

CERN

EUROPEAN ORGANIZATION FOR NUCLEAR RESEACH

A WORLDWIDE COLLABORATION

- **Founded in 1954 by 12 countries**
- **Fundamental physics research**
- **Today 22 members states**
- **World-wide collaborations**



Observers

India	220
Japan	244
Russia	982
Turkey	146
USA	979

2571

Other States

Afghanistan	1	El Salvador	1	Pakistan	41
Albania	2	Estonia	16	Palestine (O.T.).	4
Algeria	8	Georgia	36	Peru	8
Argentina	11	Gibraltar	1	Philippines	1
Armenia	25	Hong Kong	1	Saudi Arabia	3
Australia	25	Iceland	4	Senegal	1
Azerbaijan	8	Indonesia	1	Singapore	2
Bangladesh	4	Iran	28	Sint Maarten	2
Belarus	47	Ireland	22	Slovenia	27
Bolivia	3	Jordan	2	South Africa	16
Bosnia &		Kenya	1	Sri Lanka	5
Herzegovina	1	Korea, D.P.R.	1	Syria	2
Brazil	108	Korea Rep.	117	Thailand	12
Cameroon	1	Kuwait	1	T.F.Y.R.O.M.	1
Canada	134	Lebanon	12	Tunisia	6
Cape Verde	1	Lithuania	19	Ukraine	55
Chile	12	Luxembourg	4	Uzbekistan	4
China	280	Madagascar	4	Venezuela	9
China (Tapei)	45	Malaysia	15	Viet Nam	9
Colombia	30	Mauritius	1	Zimbabwe	2
Croatia	35	Mexico	64		
Cuba	7	Montenegro	3		
Cyprus	16	Morocco	12		
Ecuador	3	Nepal	5		
Egypt	19	New Zealand	7		

1415

A World-Wide Collaboration

Member States

Austria	99	Greece	152	Slovakia	88
Belgium	106	Hungary	68	Spain	337
Bulgaria	75	Israel	51	Sweden	75
Czech Republic	202	Italy	1686	Switzerland	180
Denmark	53	Netherlands	153	United Kingdom	640
Finland	87	Norway	61		
France	751	Poland	229		
Germany	1150	Portugal	109		

6352

Candidate for Accession

Romania	118

Associate Members in the Pre-stage to Membership

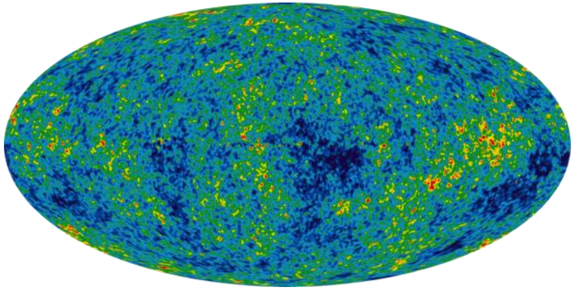
Serbia	41

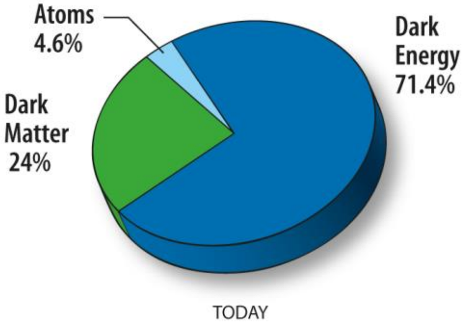
Distribution of All CERN Users by Nationality on 14 January 2014

WHAT IS THE UNIVERSE MADE OF?
HOW DIT IT START?
FUNDAMENTAL RESEARCH

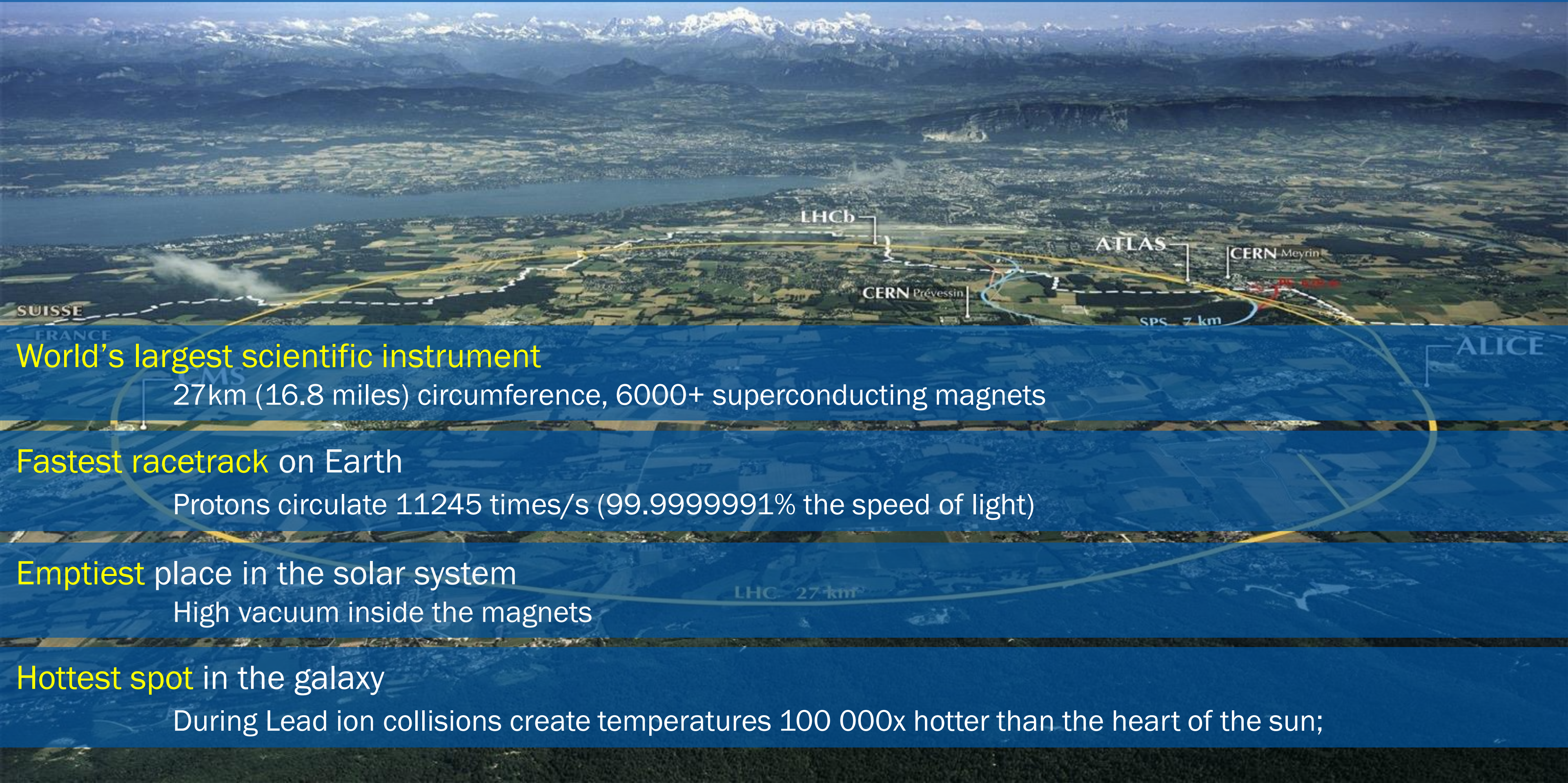
FUNDAMENTAL RESEARCH

WHY DO PARTICLES HAVE MASS?
WHY THERE IS NO ANTIMATTER LEFT?
WHAT IS 95% OF THE UNIVERSE MADE OF?





CERN Aerial View



World’s largest scientific instrument

27km (16.8 miles) circumference, 6000+ superconducting magnets

Fastest racetrack on Earth

Protons circulate 11245 times/s (99.9999991% the speed of light)

Emptiest place in the solar system

High vacuum inside the magnets

Hottest spot in the galaxy

During Lead ion collisions create temperatures 100 000x hotter than the heart of the sun;

CMS Detector

150 Million of sensor

Control and detection sensors

Massive 3D camera

Capturing 40+ million collisions per second

CERN’s INDUSTRIAL-IOT

INDUSTRIAL CONTROL SYSTEM

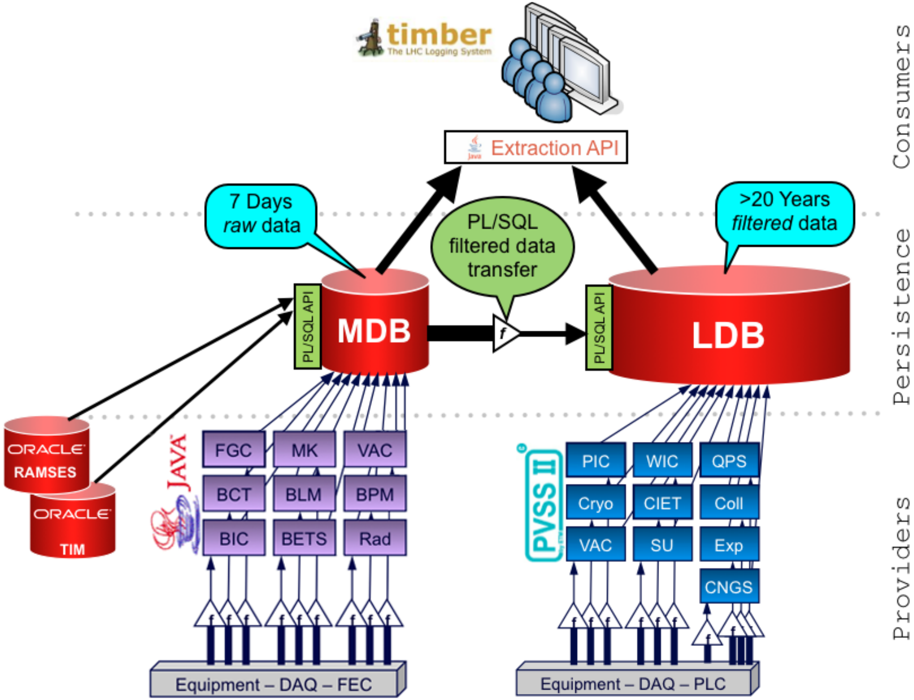
CERN ACCELERATOR LOGGING SERVICE - CALS

- In addition of physics data, CERN's produces high volume of data for its **Supervisory Control And Data Acquisition** systems.
- The scope is very wide:
  - **Accelerator systems:** cryogenics, vacuum, machine Protection, radiations...
  - **Detector Control System:** ATLAS, CMS, ALICE and LHCb
  - **Technical Infrastructure**: electrical network, cooling and ventilation systems

CERN ACCELERATOR LOGGING SERVICE - CALS

- **2 057 960 signals** produce more than **2.5TB data per day.**
- Signals range from scalars to arrays of **up-to 4 million elements.**
- **Data diverse in nature:** accelerator running modes, equipment statuses, magnet currents, cryogenics temperatures, particle beam positions, etc.
- More than **1000 individuals** and **130 expert applications**
- The system is highly tuned in terms of making use of **Oracle database features** and Oracle-specific JDBC configurations.

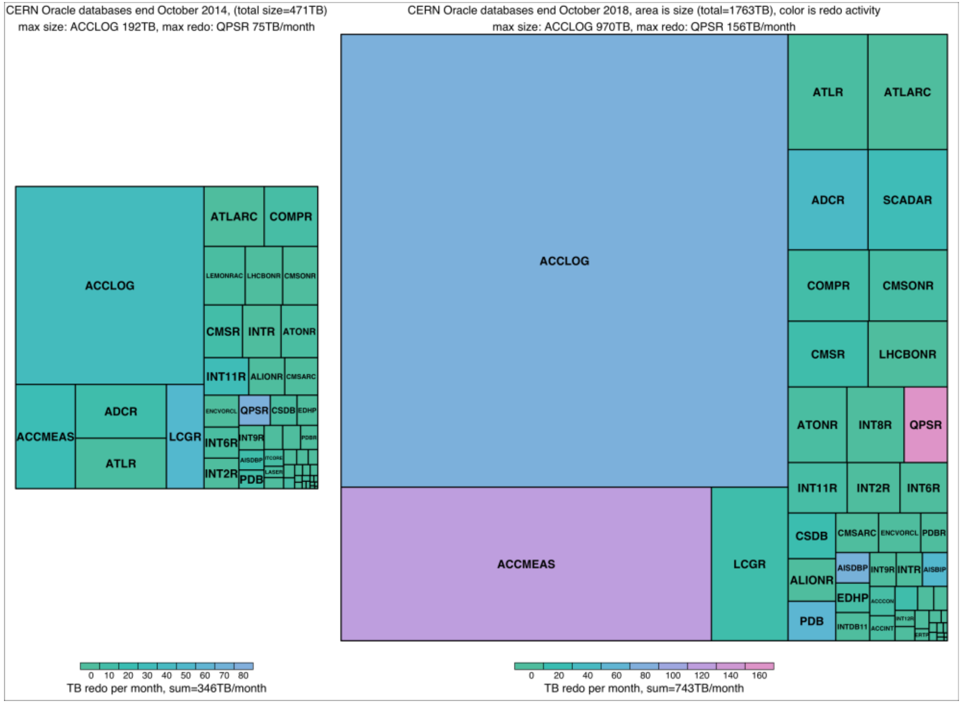
CERN ACCELERATOR LOGGING SERVICE - CALS



Credit: BE-CO-DS

CERN ACCELERATOR LOGGING SERVICE - CALS

- Compressed IIoT:
  - **1.1PB of IIoT data**
  - **2.5 TB day**
- Most active in redo:
  - **156TB/month**



CERN Oracle databases end October 2014, (total size=471TB)
max size: ACCLOG 192TB, max redo: QPSR 75TB/month

CERN Oracle databases end October 2018, area is size (total=1763TB), color is redo activity
max size: ACCLOG 970TB, max redo: QPSR 156TB/month

CERN ACCELERATOR LOGGING SERVICE - CALS

- **Advantages**
  - Simple architecture
  - Extremely efficient for 90% of use cases
  - Allow to control critical systems in almost real-time
- **Disadvantages**
  - Data exploration
  - Better performance on bigger datasets

NXCALS: Next-Generation CALS

- **Advantages**
  - Allow data exploration
  - Cost-effective solution in terms of storage
- **Disadvantages**
  - Complex set technologies
  - Increase the operation and dev costs
  - Not designed to handle critical systems

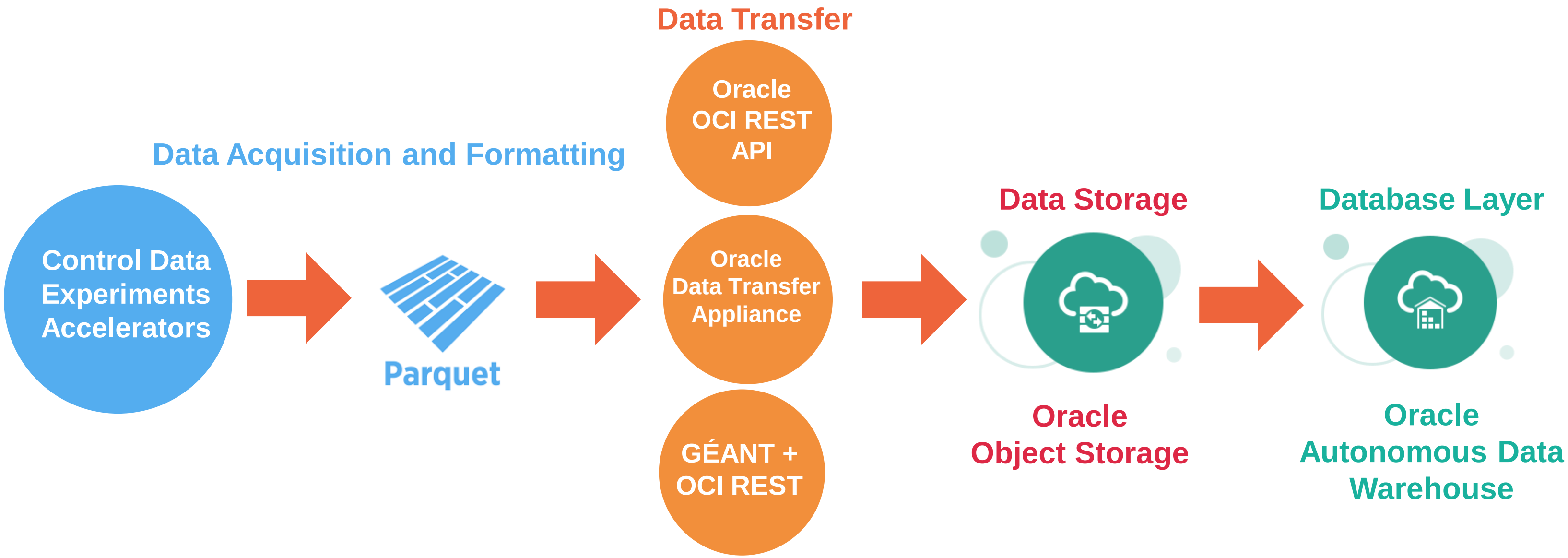
AUTONOMOUS DATA WAREHOUSE (ADW)

WHY AND HOW

ADW - WHY

- Unique system that
  - Introduces a **simple architecture**
  - Facilitates **data exploration** (unknowns – unknowns)
  - Allows to control critical systems on **almost real-time**
  - **Lowers operations and development costs**
  - **Reduces migration and integration efforts**
  - Transparent and seamless **access to advance optimization features**

ADW – HOW - GENERAL OVERVIEW



Data Acquisition and Formatting

Control Data
Experiments
Accelerators

Parquet

Data Transfer

Oracle
OCI REST
API

Oracle
Data Transfer
Appliance

GÉANT +
OCI REST

Data Storage

Oracle
Object Storage

Database Layer

Oracle
Autonomous Data
Warehouse

ADW – HOW – DATA ACQUISITION AND FORMAT

- Data is collected from the control system using **Apache Kafka** and later on is transformed into **Apache Parquet files** which are persisted in **HDFS.**
- **Parquet schema is defined on-write** based on control device categories (CMW, PVSS) and properties. Schema is used by Oracle Autonomous to **automatically generate tables definitions.**

```
root
|-- __sys_nxcals_system_id__: long (nullable = true)
|-- __sys_nxcals_entity_id__: long (nullable = true)
|-- __sys_nxcals_partition_id__: long (nullable = true)
|-- __sys_nxcals_schema_id__: long (nullable = true)
|-- __sys_nxcals_timestamp__: long (nullable = true)
|-- application_arcgroup: string (nullable = true)
|-- timestamp: long (nullable = true)
|-- value: double (nullable = true)
|-- variable_name: string (nullable = true)
```

- The data is also **partitioned by timestamp and device family.**

ADW – HOW – DATA TRANSFER

- Due to the large data volume involved **(about 1PB)** different solutions to transfer the data to **Oracle object storage** are being used:
  - **Oracle OCI Rest API**
    - On top of the OCI rest we have created a **set of scripts to automatize** the data transfer and optimize the network resources.
    - **Scan HDFS** and **bulk upload parquet file to Oracle object storage**

```
for f in $(find /hdfs/... -mindepth 2 -type d | sort -V); do
  month = $(drname $f | cut -d/ -f12)
  day = $(basename $f)
  oci os object bulk-upload -ns tenant -bn oss --src-dir $f --object
    --prefix ${entity_name}_${device}_${year}_${month}_${day}_
done
```

ADW – HOW – DATA TRANSFER

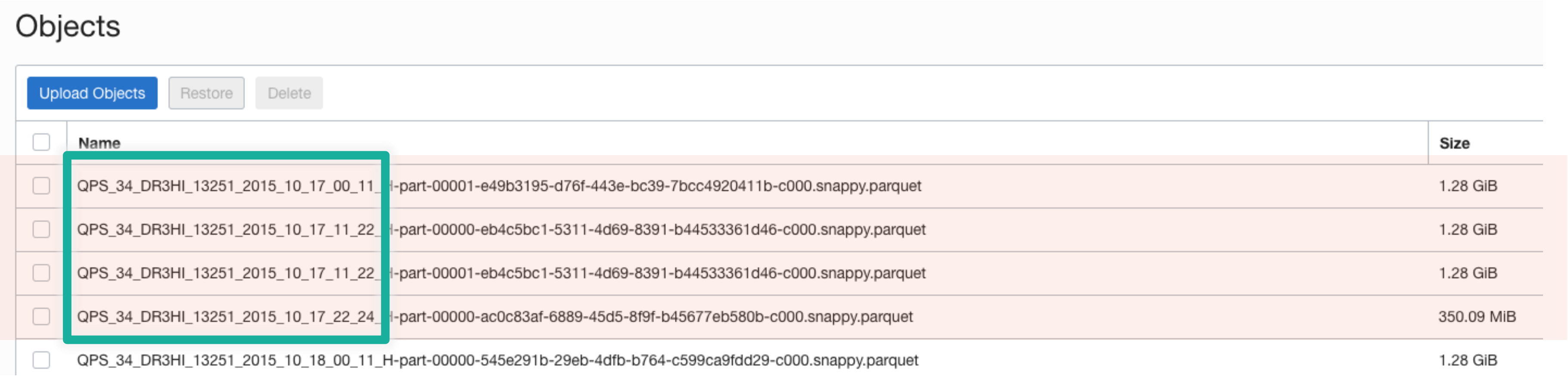
- **GEANT + OCI Rest API** – Openlab team has worked with GEANT and Oracle to make GEANT available as a provider on Oracle Cloud.
  - **BM** server **8Gbps, VMs 6Gbps**

Available Network: ***10 Gbit -> 1 GB / sec***
30 – 50% *network overhead*
Transfer time for 1 TB -> ***1000 sec / 60 = 150 min -> 2.5 hrs***
Transfer time for 1 PB -> ***2500 hrs / 24 -> 105 days -> 2.5 months***

- **Oracle Data Transfer Appliance – 150TB per machine** can be shipped to Frankfurt data center. Internal procedure need to be followed.

ADW – HOW – OBJECT STORAGE

- Data is persisted in object storage following a **specific naming logic** that is **used to create necessary partitions**



Objects

□	Name	Size
□	QPS_34_DR3HI_13251_2015_10_17_00_11_H-part-00001-e49b3195-d76f-443e-bc39-7bcc4920411b-c000.snappy.parquet	1.28 GiB
□	QPS_34_DR3HI_13251_2015_10_17_11_22_H-part-00000-eb4c5bc1-5311-4d69-8391-b44533361d46-c000.snappy.parquet	1.28 GiB
□	QPS_34_DR3HI_13251_2015_10_17_11_22_H-part-00001-eb4c5bc1-5311-4d69-8391-b44533361d46-c000.snappy.parquet	1.28 GiB
□	QPS_34_DR3HI_13251_2015_10_17_22_24_H-part-00000-ac0c83af-6889-45d5-8f9f-b45677eb580b-c000.snappy.parquet	350.09 MiB
□	QPS_34_DR3HI_13251_2015_10_18_00_11_H-part-00000-545e291b-29eb-4dfb-b764-c599ca9fdd29-c000.snappy.parquet	1.28 GiB

ADW – HOW – OBJECT STORAGE

- Object storage is scanned using Oracle **dbms_cloud packages** to determine the data have been successfully imported and **create tables and partitions by family and date**

```
FOR objects IN (
    SELECT object_name,
           substr(object_name, instr(object_name,'_',1,2)+1,
            - (instr(object_name,'_',1,5)-instr(object_name,'_',1,2)-1)) AS s_date
      FROM table(dbms_cloud.list_objects(credential_name => 'ADW_CRED_OCI_OS',
               location_uri => 'https://swiftobjectstorage.eu-frankfurt-1…’
    )) WHERE object_name LIKE ’QPS%' ORDER BY to_date(s_date, 'YYYY_MM_DD')
)LOOP
```

```
DBMS_CLOUD.CREATE_EXTERNAL_PART_TABLE(
  table_name =>'PSEN_TA',
  credential_name =>'ADW_CERD_OCI_OS',
  partition_clause => 'PARTITION BY RANGE(timestamp) (' || partition_clause || ')',
  format => json_object('type' VALUE 'parquet')
);
```

ADW – HOW – DATA MODEL

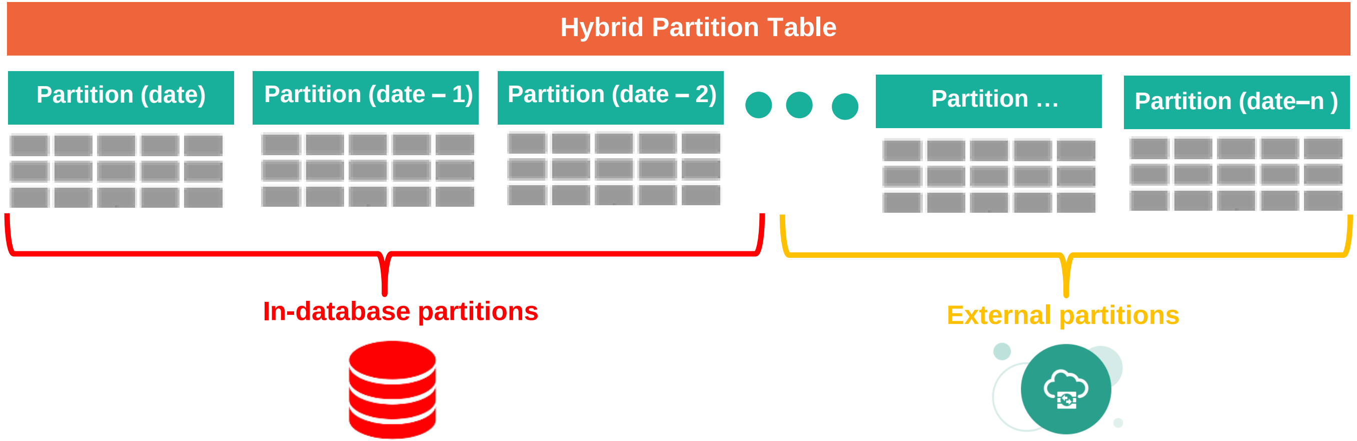
- We worked together with Oracle development and management team to define the best data model strategy:

  - Profit **from flexibility, cost-effectiveness of object storage**
  - **Reduce associated costs**
  - **Improve efficiency and performance**.

- A **hybrid rolling partition table model** based on a completed new features was implemented, tested and applied to our use case.

ADW – HOW – DATA MODEL

- The rolling hybrid model emphasizes the benefits of the oracle object storage using **transparently and coordinately:**
  - **External partitions** based on parquet files for **less accessed data** and
  - **Regular database partitions** for **data that require almost real-time responses.**



CONCLUSIONS

MOVING TO AUTONOMOUS TECHNOLOGIES

ADW – CONCLUSIONS & LESSONs LEARNT

Simplified Architecture

Automatic access to **Oracle optimization features.**

Transparent and automatic **backups and patching.**

Transparent **scale-up and scale-down** to adapt to the needs.

Brings Exadata features in a dedicated **fully managed environment**.

Operations require little or **no prior DBA expertise**.

Fully managed system **allows to focus on analytics.**

CERN

www.cern.ch