Bigdatastack.eu

@bigdatastackeu

BigDataStack

Holistic stack for big data applications and operations

Data Skipping technology

Yosef Moatti
IBM Haifa Research Labs

moatti@il.ibm.com

May the 14th – 20

Co-funded by the European Commission
Horizon 2020 - Grant # 779747

European Commission

Data Skipping: technology context

1. Big (semi) structured data
2. Object Storage
3. Apache Spark: one of the most popular Analytics engines for big data
4. More specifically Apache Spark SQL

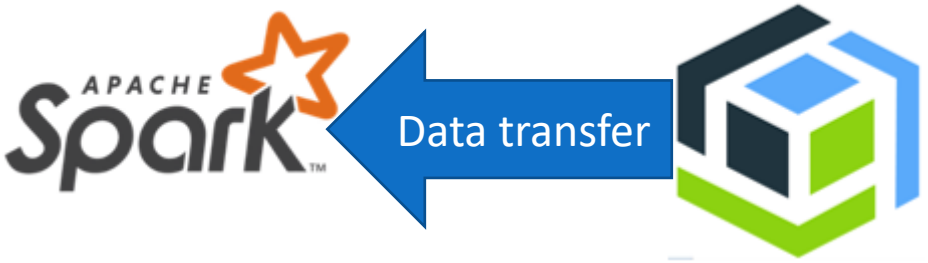
Contribution to BigDataStack requirements

- Modern BigData architectures decouple storage from compute
- Hence:
  Big data analytics implies a (big) data transfer
- Therefore:
  Data Skipping is important for BigDataStack

- *Proof*: Data Skipping is already incorporated into IBM products (see status slide)

Data Skipping: goal

1. Big (semi) structured data
2. Object Storage
3. Apache Spark: one of the most popular Analytics engines for big data
4. More specifically Apache Spark SQL



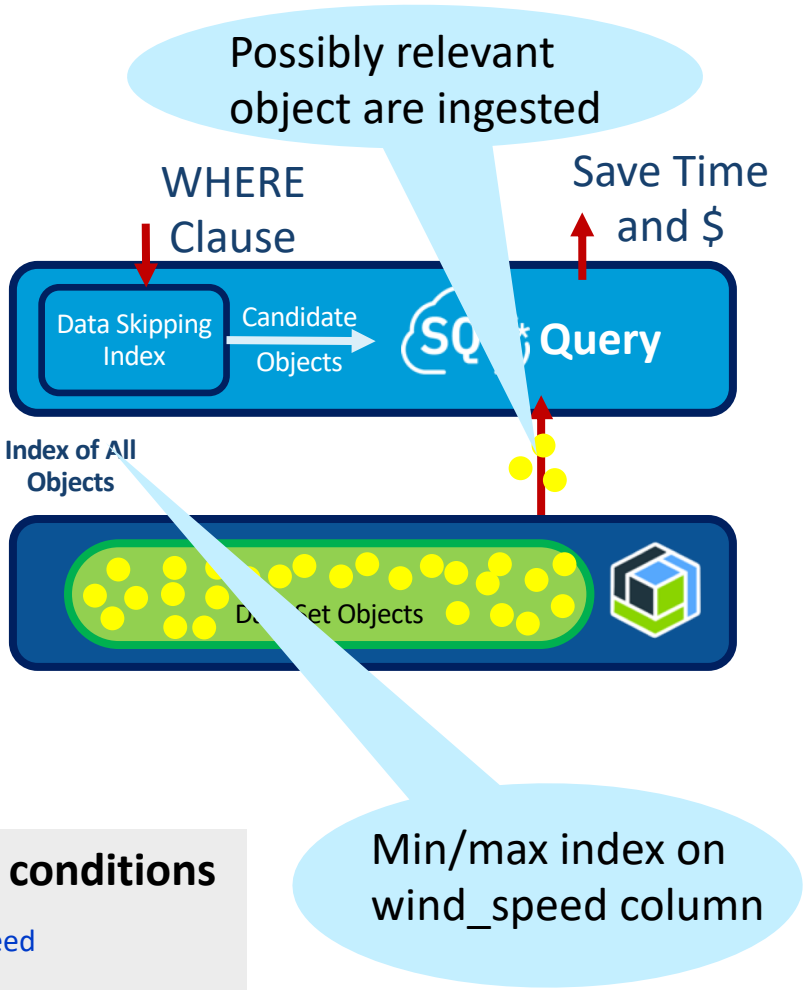
Data Skipping goal is to minimize the data transfer

Spark SQL

Data Skipping: how does it work?

- Determine which objects are NOT relevant for a SQL query using a *data skipping index*
  - Stores summary index metadata for each object
  - No change to Apache Spark base code
  - "just" take advantage of a Spark external API which permits to refine the set of objects relevant to a given query.
- Skip over irrelevant objects
  - Skipped objects are not touched at all
- IBM Data skipping is state of the art. It comprises:
  - UDFs support
  - Data Skipping for LIKE and general regular expressions
  - Multiple indexes including Bloom Filter, etc…
  - Plug-in user indexes
  - Etc…

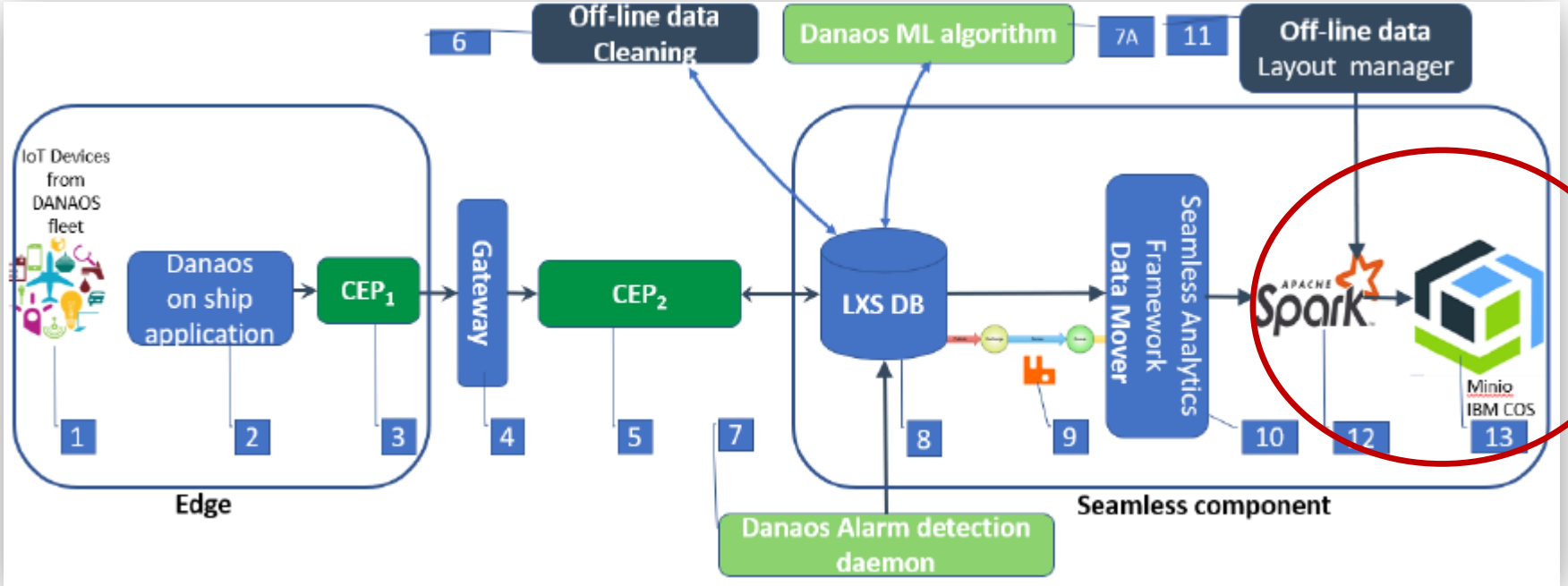
➔ Saves time and $



Example: Look for data in violent storm conditions

SELECT vessel_code, datetime, longitude, latitude, wind_speed
FROM cos://us-south/…/danaos stored as parquet
WHERE wind_speed > 30

Data Skipping within the maritime use case



- Danaos vessel dataset
- Typical SQL queries: **SELECT** vessel_code, datetime, longitude, latitude, wind_speed
  FROM cos://us-south/…/danaos stored as parquet **WHERE** wind_speed > 30
- The Object Storage: either remote IBM Cloud Object Storage (COS) (as shown) or local
- Depending on query and data partitioning we get up to 99% reduction in data transfer

Data Skipping: status

Data Skipping for the BigDataStack maritime dataset demonstrated at THINK ’19

(joint work with Danaos: the BigDatastack maritime use-case partner)

IBM Products

- Database Catalog Support added to IBM Cloud SQL Query
  - Blog published here
  - Released as **open beta**
- Data Skipping integrated within IBM Cloud SQL Query
  - Blog published here
  - Released as closed beta (as of now)
- IBM Analytics Engine (IAE)
  - Data Skipping feature available as open beta

That’s all … for now

Scientific paper soon to be submitted

Data Skipping: next steps

- Data partitioning
- Push for additional adoption
- Decouple Data Skipping from Apache Spark
- …

The team

Team:

Oshrit Feder

Guy Khazma

Gal Lushi

Yosef Moatti

Paula Ta-Shma

Technical contact point: PAULA@il.ibm.com