Studies of Big Data metadata segmentation between relational and non-relational databases

M.Golosova (NRC KI), G.Dimitrov (CERN), M.Grigorieva (NRC KI),
A.Klimentov (BNL), M.Potekhin (BNL), E.Ryabinkin (NRC KI)

National Research Centre «Kurchatov Institute»
CERN
Brookhaven National Laboratory

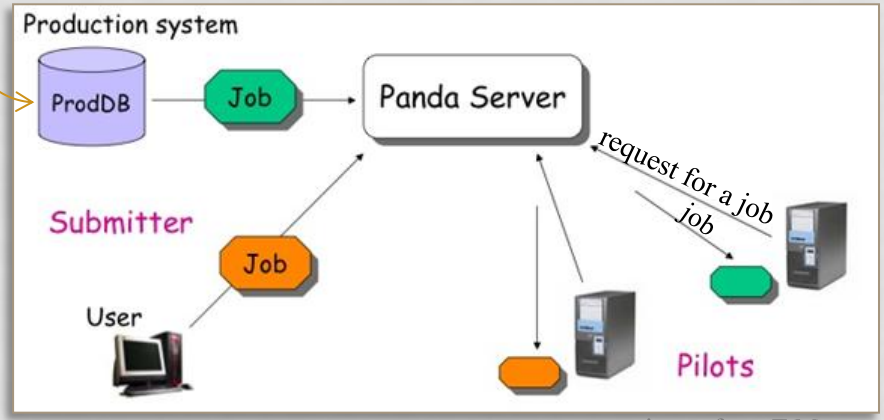
CHEP2015, Okinawa, Japan, April 13-17 2015

Outline

- **PanDA WMS**
  - Metadata
  - Monitoring
  - Data specification
- **NoSQL as archive**
  - Precalculation for speedup
- **Study**
  - Test set-up
  - Scaling
  - Precalculation effect
  - NoSQL archive performance

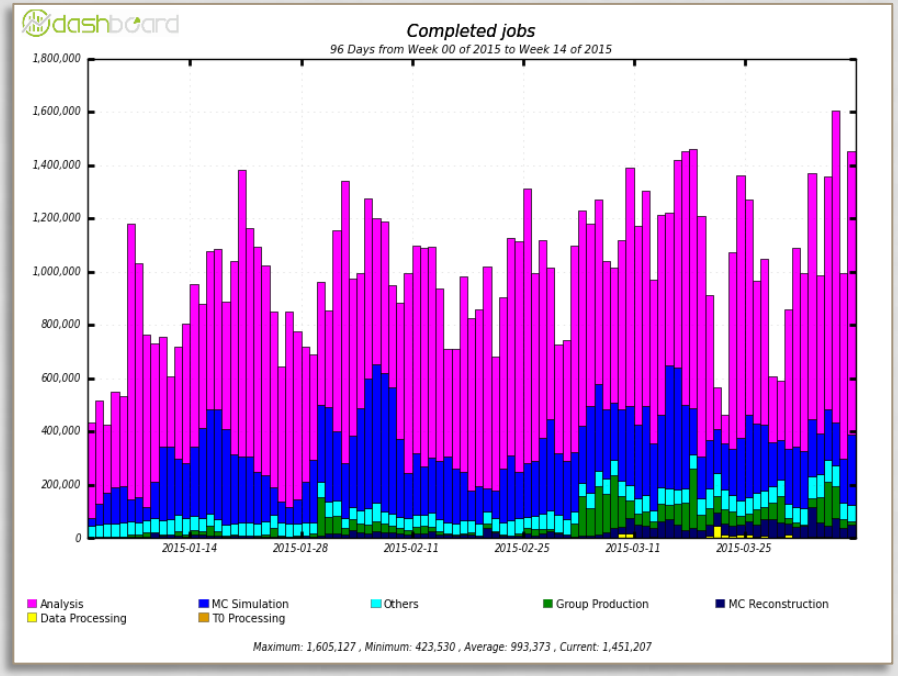
PanDA WMS

Metadata storage

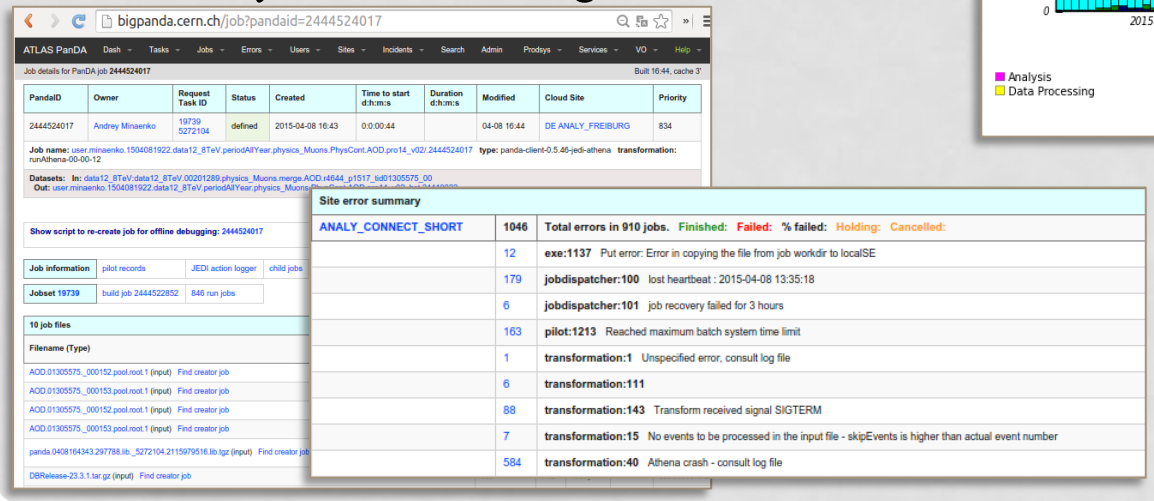


courtesy picture from T.Maeno

Scale of the challenge: up to 1.7M completed jobs per day



Metadata analysis and monitoring: PanDA Monitor

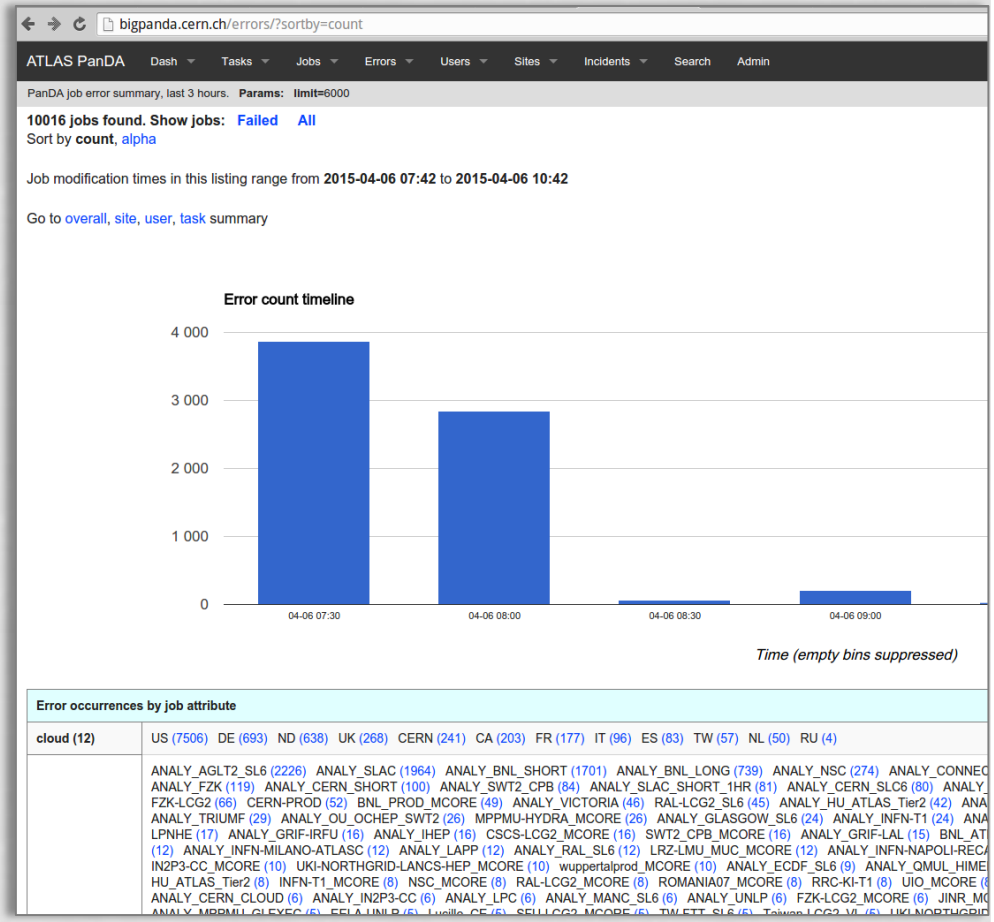


Related CHEP talks: T.Maeno, [The Future of PanDA in ATLAS Distributed Computing](#) (CHEP ID 144)

PanDA monitoring

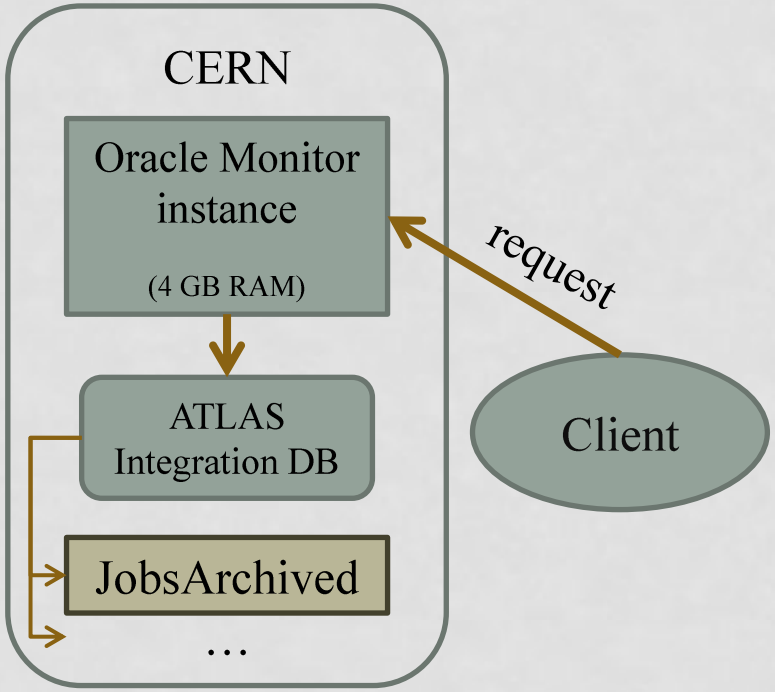
- **PanDA Monitor**
  Web-based analytical interface providing information about jobs and tasks within the system

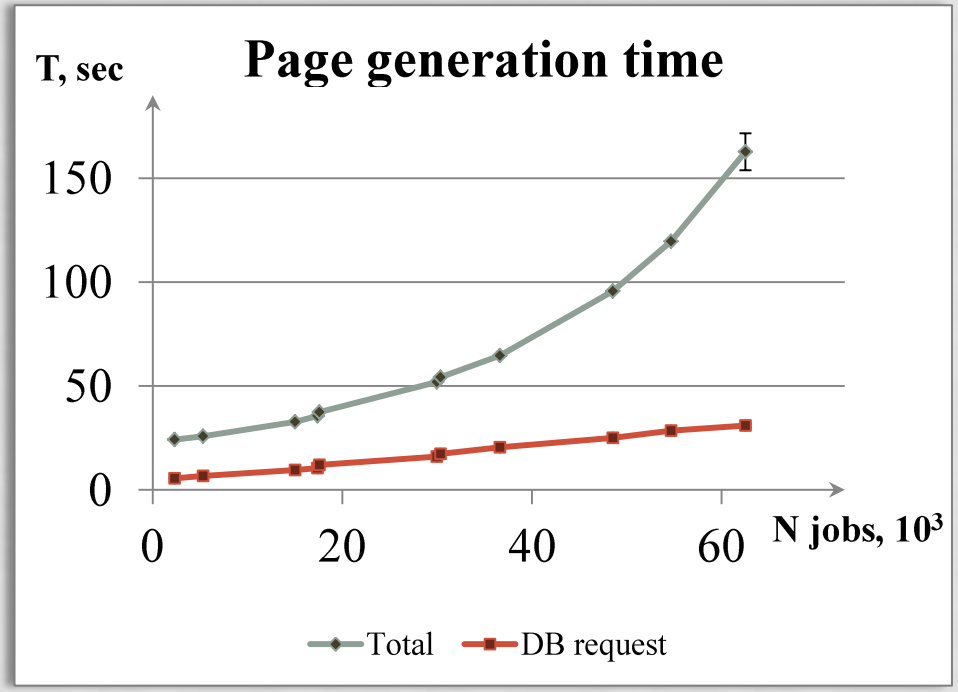
- **Page «Errors»**
  Page of PanDA Monitor providing information about job errors that have occurred in the system



http://bigpanda.cern.ch/errors/

Job errors monitoring

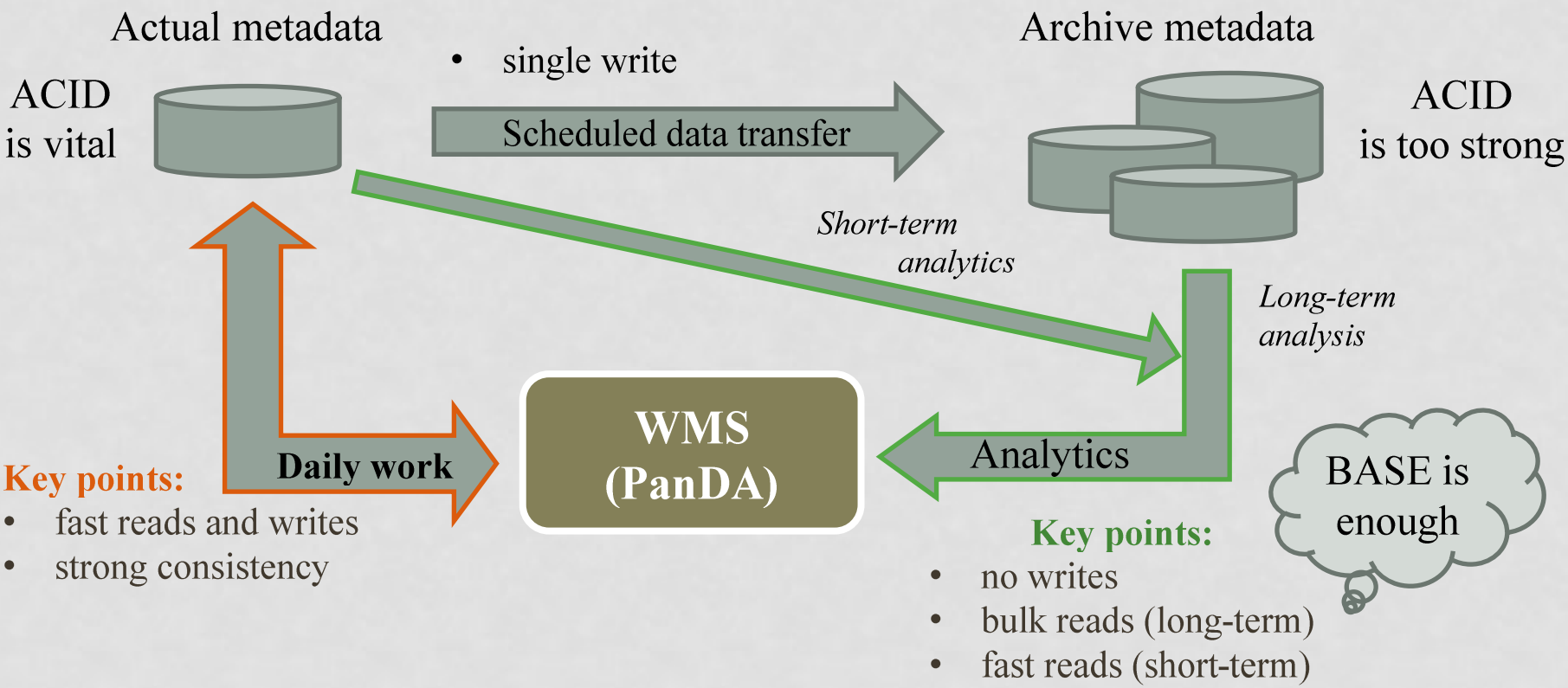




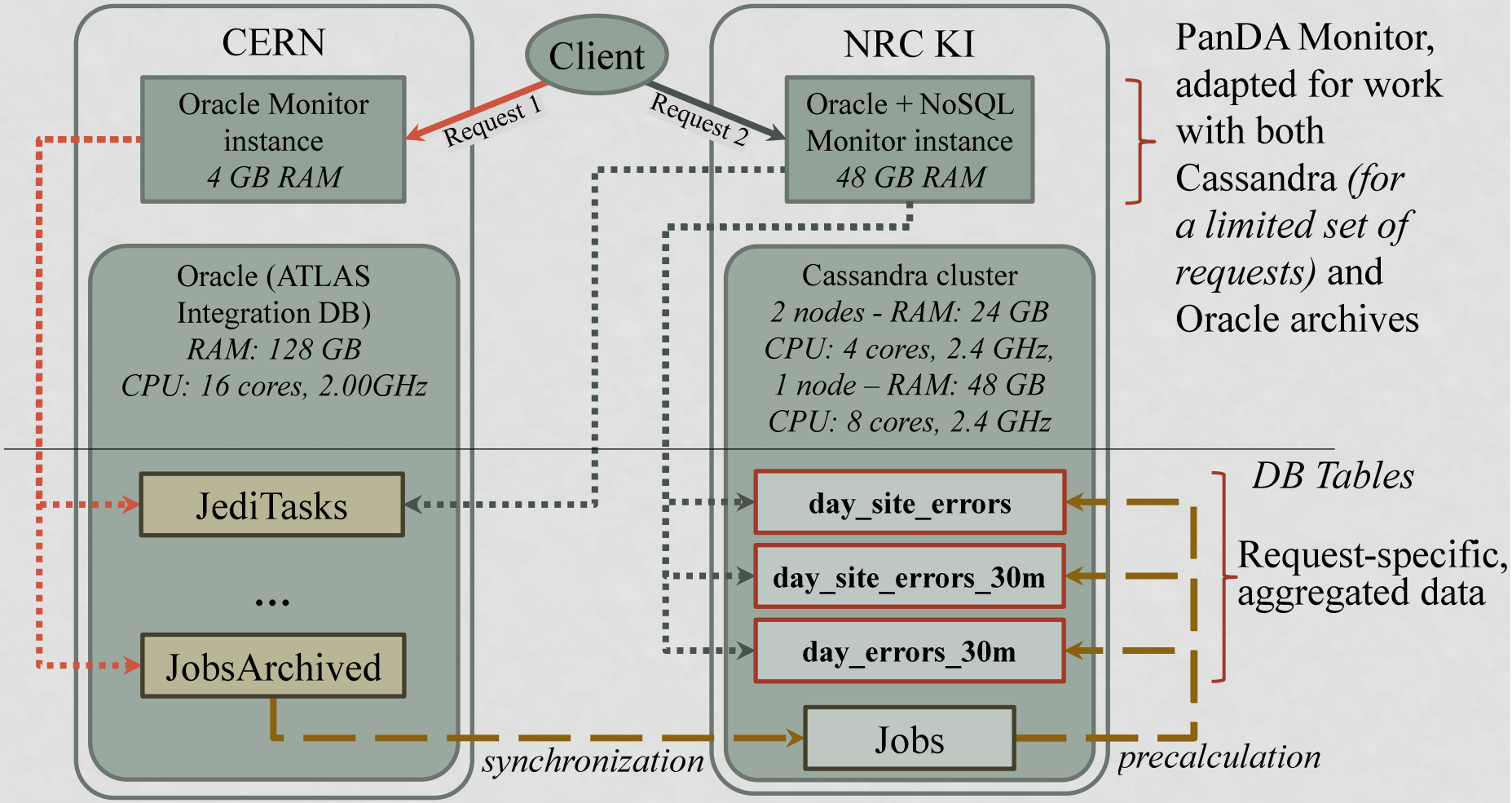
1) Get list of jobs from DB (according to request parameters) ⇨ **2) Summaries by parameters (overall, computing site, user, task)** ⇨ **3) Get additional information from DB** ⇨ **4) Generate output**

Specificity of the metadata storage

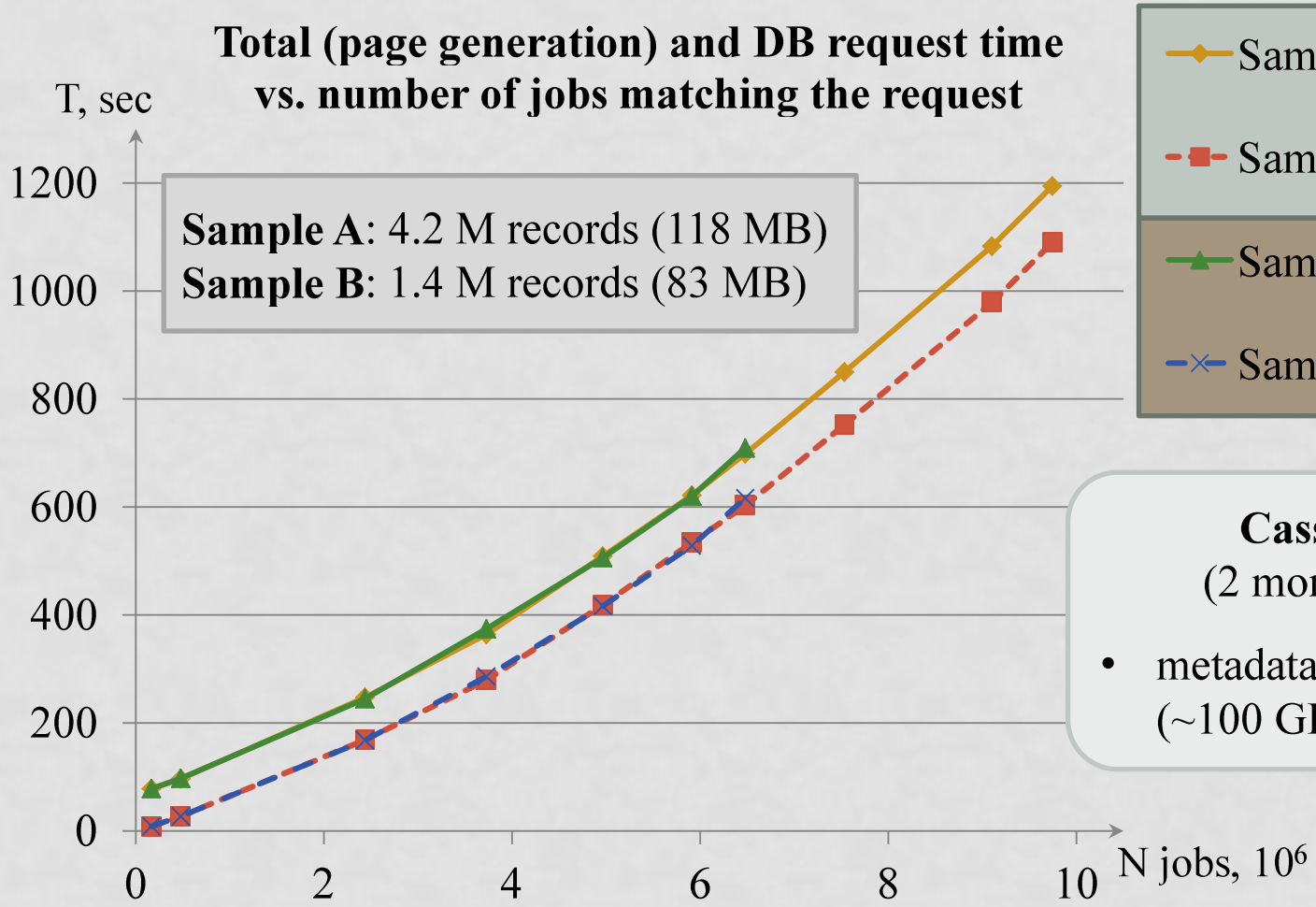
- Number of requested records (ATLAS PanDA Archive: ~900M of jobs)
- RDBMS as archive backend (ACID standard)



Test set-up


CERN
Client
NRC KI
Oracle Monitor instance
4 GB RAM
Request 1
Request 2
Oracle + NoSQL Monitor instance
48 GB RAM
PanDA Monitor, adapted for work with both Cassandra (for a limited set of requests) and Oracle archives
Oracle (ATLAS Integration DB)
RAM: 128 GB
CPU: 16 cores, 2.00GHz
Cassandra cluster
2 nodes - RAM: 24 GB
CPU: 4 cores, 2.4 GHz,
1 node – RAM: 48 GB
CPU: 8 cores, 2.4 GHz
DB Tables
JediTasks
...
JobsArchived
day_site_errors
day_site_errors_30m
day_errors_30m
Request-specific, aggregated data
Jobs
synchronization
precalculation

Cassandra scaling test



Sample A – total

Sample A – DB request

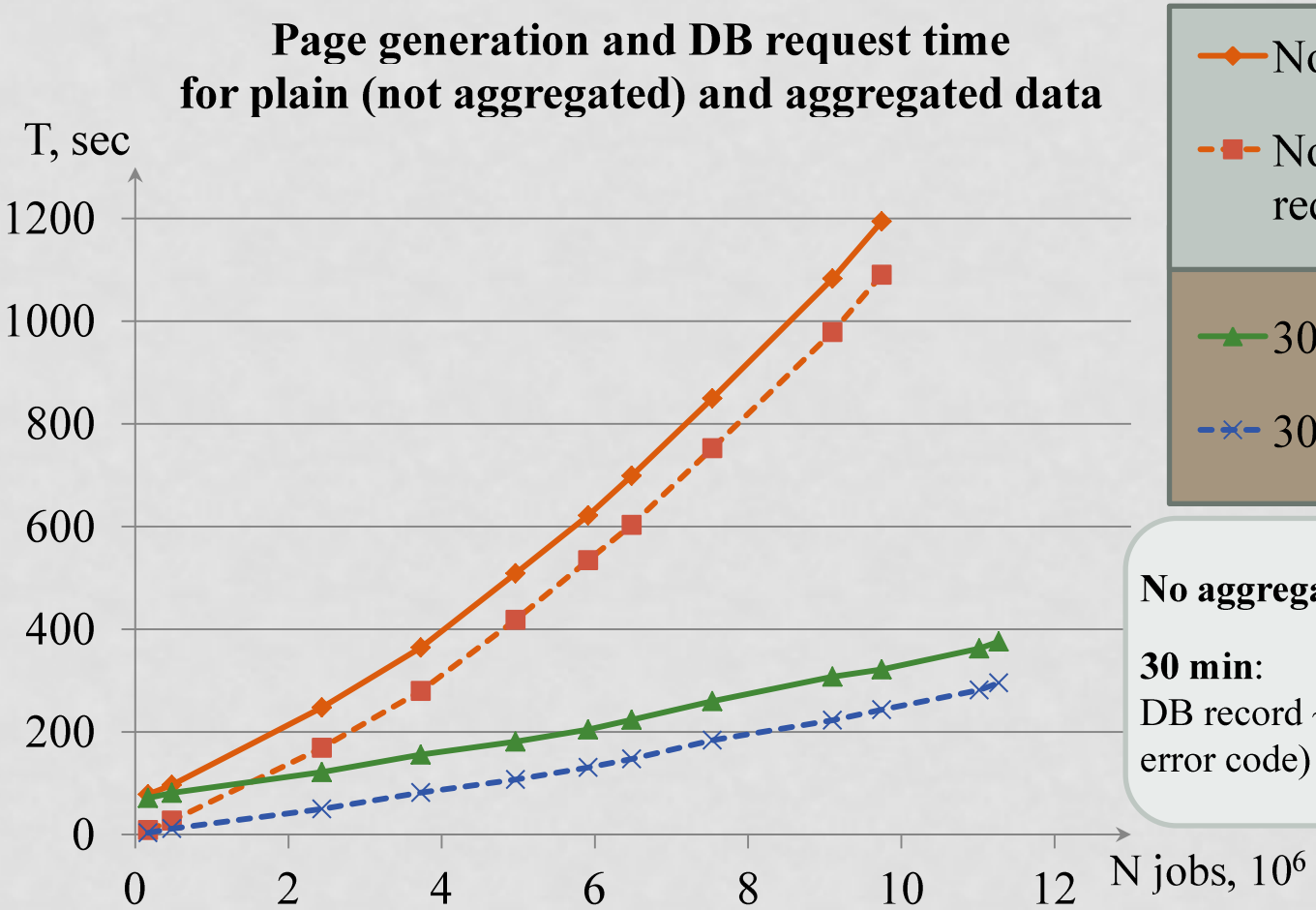
Sample B – total

Sample B – DB request

Cassandra DB
(2 months archive):

- metadata for ~60 M jobs (~100 GB per replica)

Aggregation effect test



No aggregation — total

No aggregation — DB request

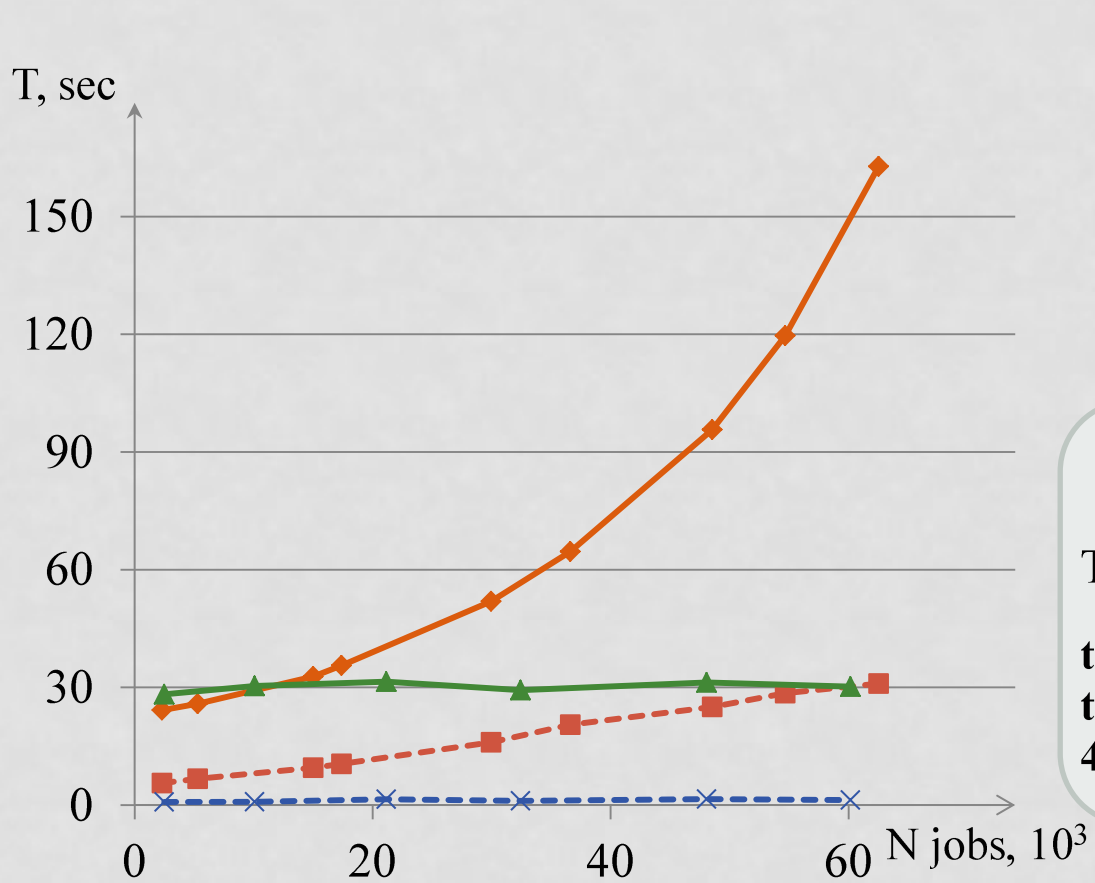
30 min — total

30 min — DB request

No aggregation: DB record ~ 1 error

30 min:
DB record ~ all errors (with the same error code) within given 30 minutes

NoSQL Archive performance



Oracle — total

Oracle — DB request

Cassandra — total

Cassandra — DB request

Values for Cassandra

$T = 4*(t - t_{add}) + t_{add}$, where:

t – total page generation time

$\mathbf{t_{add}}$ – not a summary-specific DB requests

4 – number of specific summaries

Summary

- It is hardly possible to perform long-term metadata analysis without any precalculation.
- Replacing RDBMS with NoSQL in archive part of metadata storage can be used to improve availability of historical data.
- Prototype of NoSQL (Cassandra) archive was created and tested on a 2-month slice of metadata from ATLAS PanDA Archive.
- First results look promising. NoSQL archive will be extended to confirm this belief in new tests.
- Adaptation of PanDA Monitor for work with NoSQL archive will be continued.

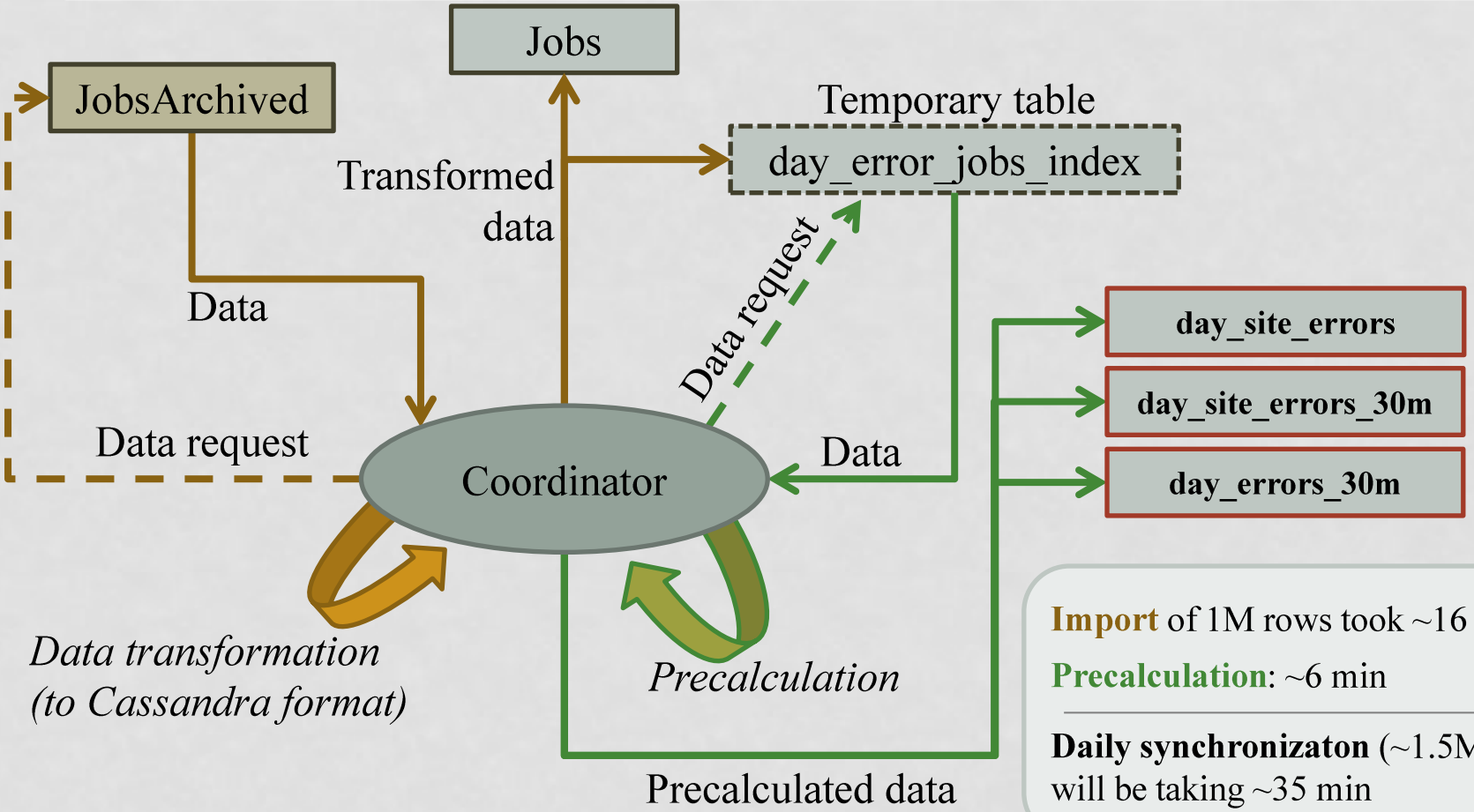
Acknowledgment

- This talk drew on presentations, discussions, comments, input from many our CERN, ATLAS, NRC-KI Colleagues, thanks to all people working in PanDA project

- This work was funded in part by the Russian Ministry of Science and Education under Contract N14.Z50.31.0024

Back up slides

Synchronization & Precalculation



Import of 1M rows took ~16 min

Precalculation: ~6 min

Daily synchronizaton (~1.5M jobs) will be taking ~35 min

Cassandra Column Families

Primary Key
- Partition Key
- Clustering Key

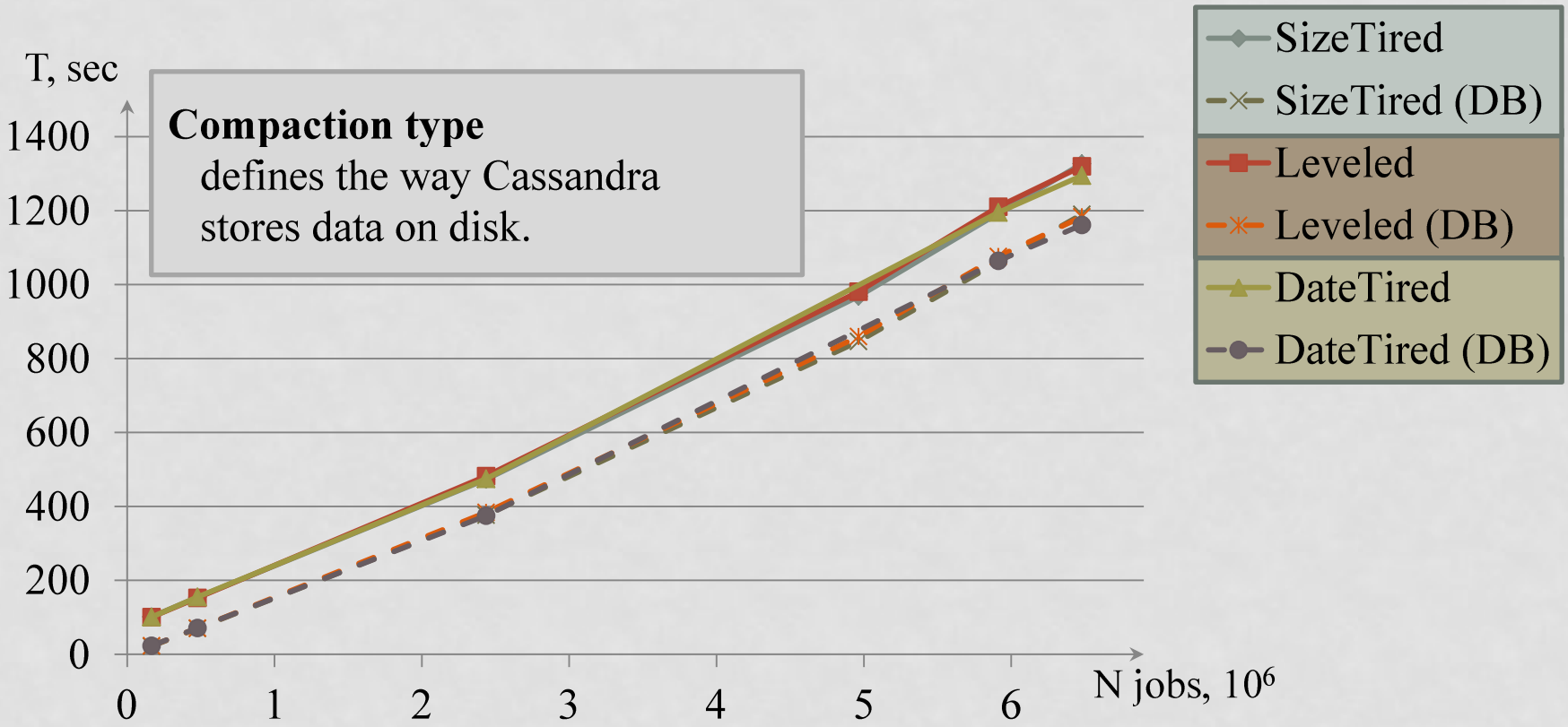
Columns
- Common
- Counter

day_site_errors_cnt_30m						
date	computingSite	base_mtime	errcode	errdiag	err_count	job_count

day_site_errors				
date	computingSite	errcode	pandaid	errdiag

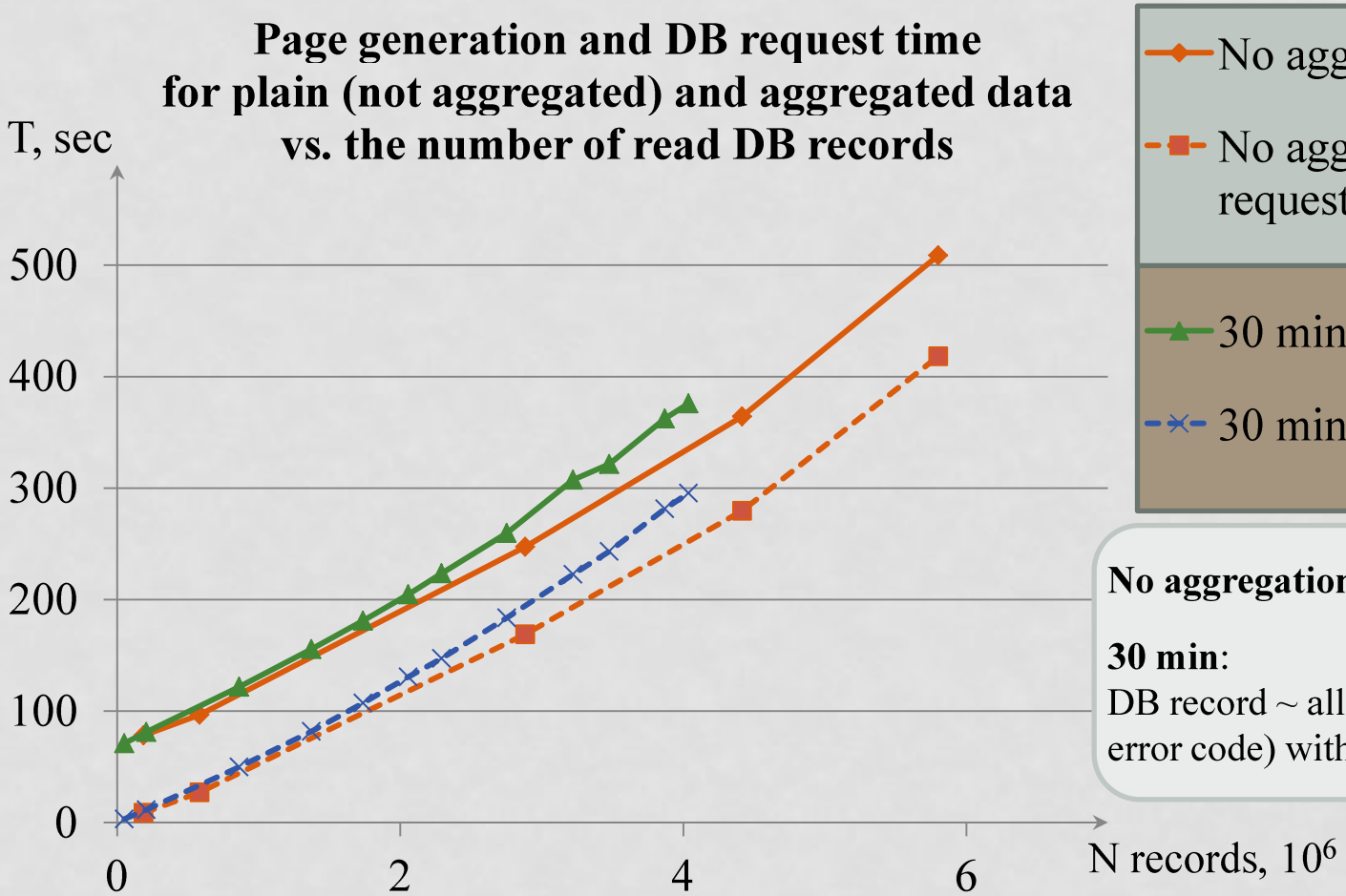
day_errors_30m		
date	base_mtime	count

Compaction types



Compaction type
defines the way Cassandra stores data on disk.

Aggregation effect test (DB records)



No aggregation — total

No aggregation — DB request

30 min — total

30 min — DB request

No aggregation: DB record ~ 1 error

30 min:
DB record ~ all errors (with the same error code) within given 30 minutes