Data warehousing on Hadoop
7 months later

Marek Grzenkowicz
Roche Polska

Agenda

- 7 months ago
- Today
- Tomorrow
- In a few months

7 months ago

2015.09.29 Data warehousing on Hadoop

http://go.roche.com/strada2015

Some context

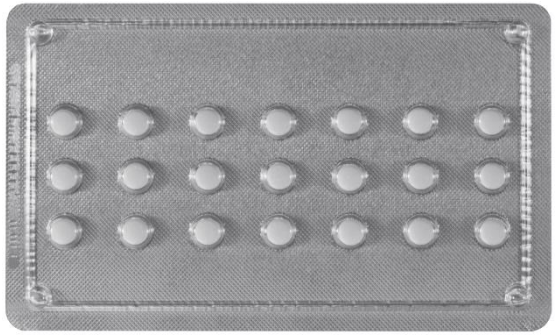
REPLAY

Roche

Roche

Roche Pharmaceuticals (Pharma)

Roche Diagnostics (Dia)



Objectives of the project StraDa

REPLAY

Roche

Measure and **compare** the performance of labs

- Workload
- Turnaround time (TAT)

Discover and **understand** the reasons

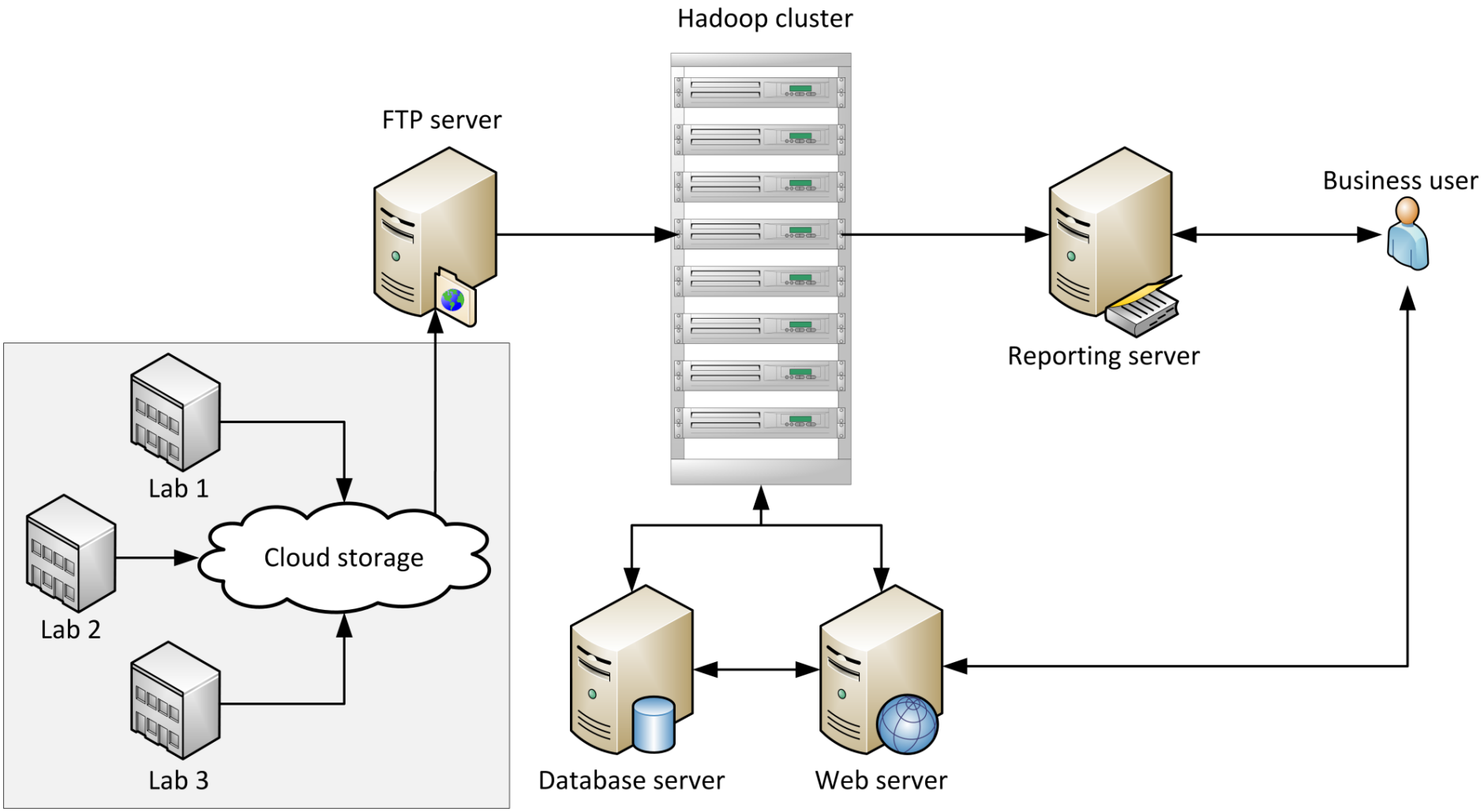
- Hardware configuration
- Tasks and work organization
- Other, unknown factors

There are **7 types** of the Workload KPI and **19** TATs.

Overview of the data flow

REPLAY

Roche



Conclusion

Roche

Can I build a data warehouse on Hadoop?	▪ Yes.
Can the star schema be used?	▪ Yes, but it is a usability/performance tradeoff. ▪ YMMV, so test it carefully.
Can I just put Hadoop in place of my RDBMS?	▪ No, it is way more complex than that.
Is it worth it?	▪ It depends, so don't follow the Big Data hype blindly.

Today

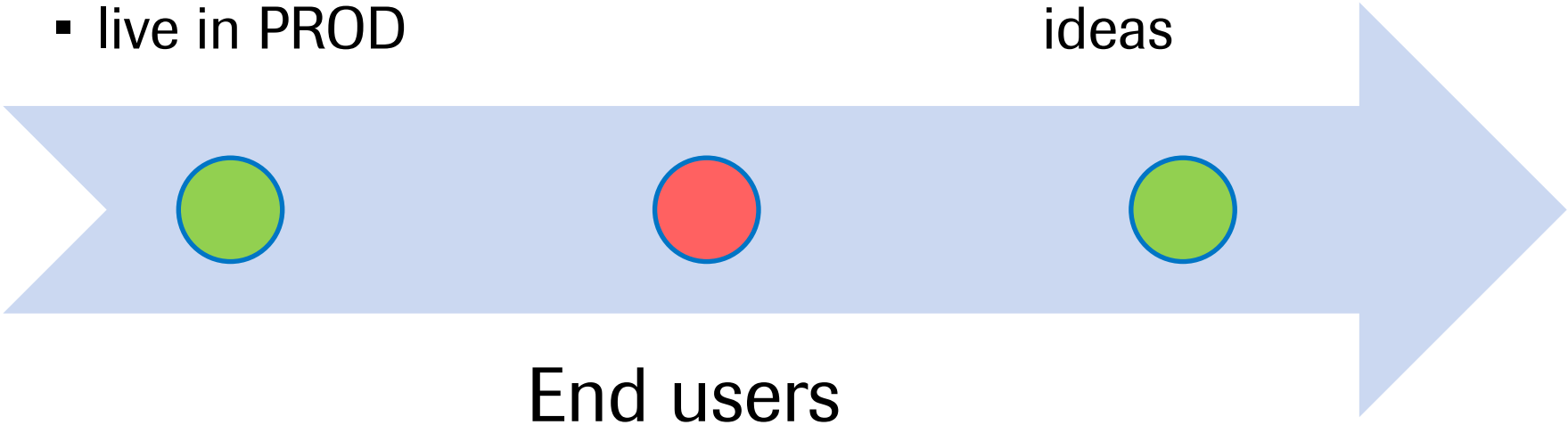
Roche

Solution

- live in PROD

Developers

- always full of ideas



End users

- unhappy about the performance

Data volume

Input files	Fact tables	
720 GB x2.5	1.1 billion (10^9) rows x3	17 GB x7
	2.2 billion (10^9) rows x1.5	87 GB x5.5

Input files: uncompressed text files; 1 country, 23 months

Fact tables: compressed Parquet files

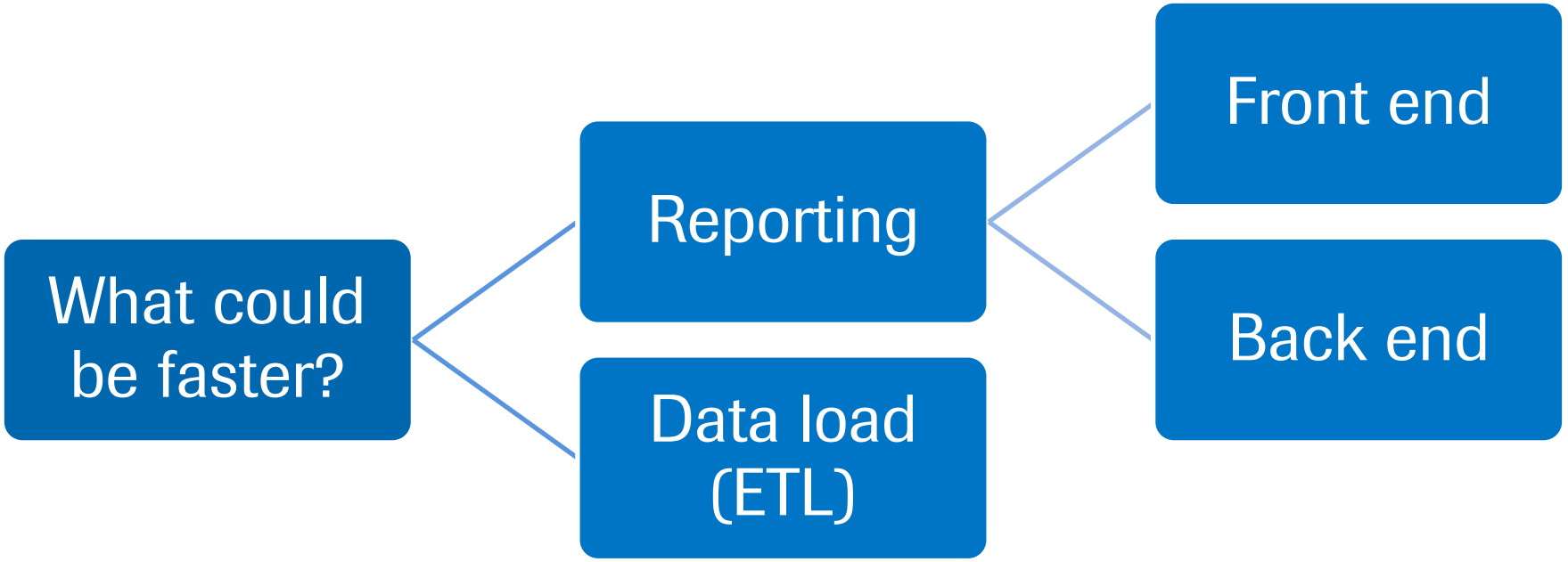
xN – N times more than in September 2015

Key facts

1. Percentile metrics (median, 95th) instead of averages
2. Multistep workflow implementing complex business logic and numerous data quality rules
3. Stringent performance requirements for the reports:

Data range	Initial load	Refresh
7 days	< 5 s	< 1 s
1 month	< 10 s	< 3 s
1 year	< 60 s	< 5 s

Performance aspects



Reporting performance – front end

- The partition pruning optimization cannot be utilized by Impala, because Tableau **filters don't support separate *Value* and *Display* aliases**
  - Filters show network names and tables are partitioned by network IDs
- Denormalizing the star schema solves the problem of inefficient JOINs but **slows down the filters**
  - Filters cannot be loaded concurrently
  - Every filter now needs to perform a full scan on a huge table to load a handful of distinct values
- Tableau generates **suboptimal queries**
  - The 3rd party connectors may be to blame for that

Reporting performance – example query

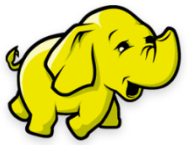
```
SELECT
COUNT(DISTINCT (CASE WHEN false THEN `ftatimelc`.`sampleid` WHEN NOT false THEN
CONCAT(CONCAT(CONCAT(CONCAT(CAST(`ftatimelc`.`sampleid` ASSTRING),';'),CAST(`ftatimelc`.`parameter_sk` AS STRING)),';')
,CAST(`ftatimelc`.`sampletype_sk` AS STRING)) ELSE CAST(NULL AS STRING) END)) AS `temp_tc___24930455`,
MIN(`ftatimelc`.`tat`) AS`temp_tc___27889298`,
SUM(1) AS `temp_tc___39677625`,
MAX(`ftatimelc`.`tat`) AS `temp_tc___44831943`,
COUNT((CASE WHEN false THEN `ftatimelc`.`sampleid` WHEN NOT false THEN
CONCAT(CONCAT(CONCAT(CONCAT(CAST(`ftatimelc`.`sampleid` AS STRING),';'),CAST(`ftatimelc`.`parameter_sk` AS STRING)),';'
),CAST(`ftatimelc`.`sampletype_sk` AS STRING)) ELSE CAST(NULL AS STRING) END)) AS `cnt_calculation_15`,
MAX((CASE WHEN (((CAST((CASE WHEN 5 = 0 THEN CAST(NULL AS DOUBLE) ELSE CAST(`ftatimelc`.`tat` AS DOUBLE) / 5
END) AS BIGINT) * 5) - (CASE WHEN (`ftatimelc`.`tat` < 0) THEN 5 WHEN NOT (`ftatimelc`.`tat` < 0) THEN 0 ELSE
CAST(NULL AS INT) END)) >= 180.) THEN 1 ELSE 0 END)) AS `max_calculation_02`

FROM `strada_db`.`ftatimelc` `ftatimelc`
JOIN `strada_db`.`security_auth` `security_auth` ON (`ftatimelc`.`customer_sk` = `security_auth`.`customer_sk`)

WHERE ((((`security_auth`.`login_name` = 'example.com\\foobar') AND(((`ftatimelc`.`month` >=
CONCAT(CONCAT(CAST(2015 AS STRING),'-'),(CASE WHEN false THEN CAST(5 AS STRING) WHEN NOT false THEN
CONCAT('0',CAST(5 AS STRING)) ELSE CAST(NULL AS STRING) END))) <> 0) AND((`ftatimelc`.`month` <=
CONCAT(CONCAT(CAST(2015 AS STRING),'-'),(CASE WHEN false THEN CAST(7 AS STRING) WHEN NOT false THEN
CONCAT('0',CAST(7 AS STRING)) ELSE CAST(NULL AS STRING) END))) <> 0))) AND((('Test' =
`ftatimelc`.`ws_workflowspanobjecttype`) <> 0) AND ((TRUNC(`ftatimelc`.`fulldate`, 'DD') >= CAST('2015-05-01
00:00:00' AS TIMESTAMP)) <> 0) AND ((((CAST((CASE WHEN 5 = 0 THEN CAST(NULL AS DOUBLE) ELSE
CAST(`ftatimelc`.`tat` AS DOUBLE) / 5 END) AS BIGINT) * 5) - (CASE WHEN (`ftatimelc`.`tat` < 0) THEN 5 WHEN NOT
(`ftatimelc`.`tat` < 0) THEN 0 ELSE CAST(NULL AS INT) END)) >= 0) AND (((CAST((CASE WHEN 5 = 0 THEN
CAST(NULL AS DOUBLE) ELSE CAST(`ftatimelc`.`tat` AS DOUBLE) / 5 END) AS BIGINT) * 5) - (CASE WHEN (`ftatimelc`.`tat` <
0) THEN 5 WHEN NOT (`ftatimelc`.`tat` < 0) THEN 0 ELSE CAST(NULL AS INT) END)) <=
800)) AND ((TRUNC(`ftatimelc`.`fulldate`, 'DD') <= CAST('2015-07-31 00:00:00' AS TIMESTAMP)) <>
0) AND (`ftatimelc`.`lc_end_labconfiguration` = 'BIOLAB - 78945612300') AND ((`ftatimelc`.`lc_end_modulefamily` >=
'Hematologie analyzers') AND (`ftatimelc`.`lc_end_modulefamily` <= 'cobas
8000')) AND (`ftatimelc`.`parameter_isincludedintatcalculation` IN('U', 'Y')) AND (`ftatimelc`.`ws_workflowspan` =
'Total Registration')))

HAVING (COUNT(1) > 0)
```

Reporting performance – back end

- It is simply **not possible to query 2 billion rows** in real time
  - And there is really no need for that – the standard reports usually require only 2-3 months worth of data
  - There are two distinct user groups with very different needs (lab consultants and data scientists) that must be addresses using different tools and methods
- Some metrics can and should be **precomputed**
  - However, percentile is **inherently sequential** - MPP cannot make it faster and appropriate SMP is prohibitively expensive
- The chosen partitioning key causes uneven distribution of data across the partitions (**data skew**)
  - Nodes that process small partitions are underutilized and the others become bottlenecks

Reporting performance – data load (ETL)

- Make things as simple as possible, **but not simpler**
  - The naive approach of storing the data in the most granular form and always processing the entire data set does not work
- MapReduce is a **disk-based** computing engine
  - For complex workflows it becomes a major bottleneck
  - Scripts developed in tools like Pig or Hive sometimes translate into a long chain of M/R jobs and there is no easy way to reduce the number of them
  - Spark can perform better; if available



Tomorrow

- Optimized **partitioning schemes** and **Parquet block size**
  - It requires an iterative approach and a lot of experimenting
  - When partition pruning can be used and the data skew is minimized, some queries may become even ~5x faster

In a few months

Roche

- **Polyglot processing architecture** (Lambda, Kappa, Iota) – different workloads handled by different tools
  - Micro-batching (to process files as they arrive)
  - Nightly delta load (to reprocess sets of files)
  - Big batch (to perform final data consolidation; once a month, possibly offloaded to the cloud with only aggregated results fetched back)
- Going **beyond Hadoop**
  - Hadoop for the batch and speed layer
  - Some other tool (e.g. an RDBMS) for the serving layer
- **Percentile calculation push-down** – moving the calculation from the front end (Tableau) to the back end (Impala)
  - Requires code development by Cloudera and/or Tableau

Summary

```
01: if usingHadoop:
02:     isAnythingPossible = True

    Traceback (most recent call last):
      File "<stdin>", line 2, in <module>
    ReasoningError: powerful does not mean omnipotent
```

Questions and (hopefully) answers

marek.grzenkowicz@roche.com

http://it.roche.pl/

Recommended materials

1. Polyglot persistence and processing
   - http://martinfowler.com/bliki/PolyglotPersistence.html
   - http://datadventures.ghost.io/2014/07/06/polyglot-processing/
2. Lambda architecture
   - https://en.wikipedia.org/wiki/Lambda_architecture
   - http://jameskinley.tumblr.com/post/37398560534/the-lambda-architecture-principles-for
   - https://www.mapr.com/developercentral/lambda-architecture
3. Kappa architecture
   - https://www.oreilly.com/ideas/questioning-the-lambda-architecture
4. Iota architecture
   - http://iot-a.info/
   - http://www.slideshare.net/Hadoop_Summit/a-modern-iot-data-processing-toolbox

Doing now what patients need next