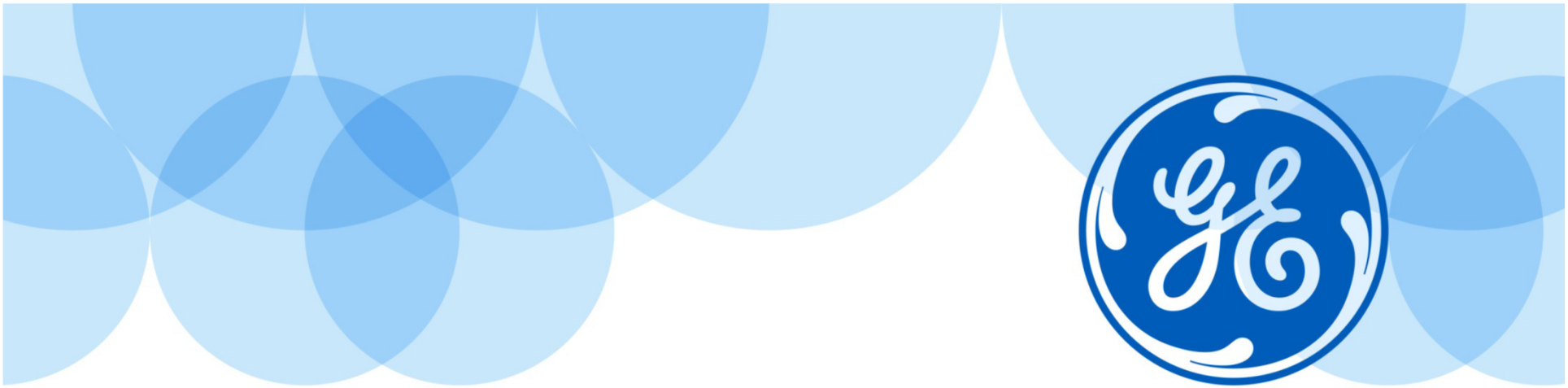


Experimenting with Lucene-based

Search engines

Stories about Fast Data

Gergely Devenyi & Akos Barabas
Budapest NoSQL Forum - March 23, 2016

Imagination at work

Agenda

1. Team intro
2. Use case highlights
   a) SolR
   b) Elasticsearch
3. Tool comparison
4. Recommended goody

Team intro

Team intro

- Who – GE Digital -> Predix Data Science
- Where – Vaci Greens -> Digital Hub
- What – Full spectrum of data analytics
- How – Data Lake; Cloud Foundry -> Predix
- Why – Power of 1%; Digital Industrials

Use case highlights

Solr

Solr – overview

Goal	provide fast data for internal GE portals
Context	sourcing data resides in GPDB warehouse GPDB architecture optimized for analytics
Problem	slow parallel retrievals multi-tenancy -> resource mgmt challenge
Solution	Lucene search & indexing technology to go for Solr endpoints adhere to REST API standards GPDB layer -> Solr index -> API endpoint (40M+ documents through 9 endpoints)

Solr – performance example

GPDB based API — GREENPLUM

Thread Stats	Avg	Stdev	Max	+/- Stdev
Latency	1.47s	366.27ms	1.96s	63.16%
Req/Sec	4.97s	6.85	50.00	91.59%

Latency Distribution	50%	75%	90%	99%
	1.61s	1.81s	1.96s	1.96s

1150 requests in 1.00m, 843.41KB read

Requests/sec: 19.13

Transfer/sec: 14.03KB

Solr based API — SOLR

Thread Stats	Avg	Stdev	Max	+/- Stdev
Latency	32.35ms	46.37ms	347.59ms	77.95%
Req/Sec	0.96k	1.02k	3.46k	74.97%

Latency Distribution	50%	75%	90%	99%
	4.09ms	66.46ms	104.74ms	179.12ms

450402 requests in 1.00m, 584.76MB read

Requests/sec: 7497.01

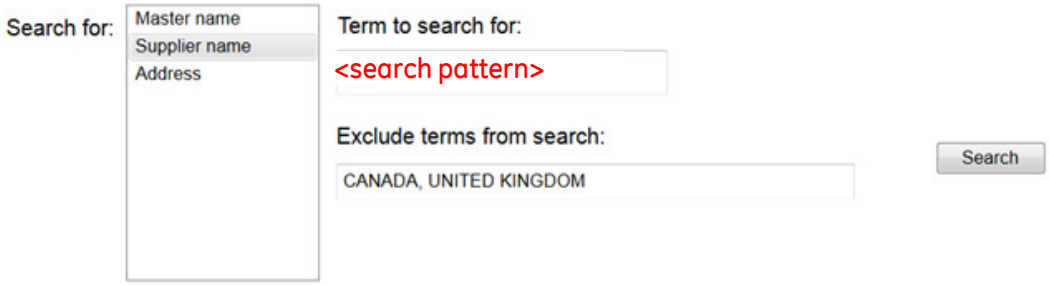
Transfer/sec: 9.73MB

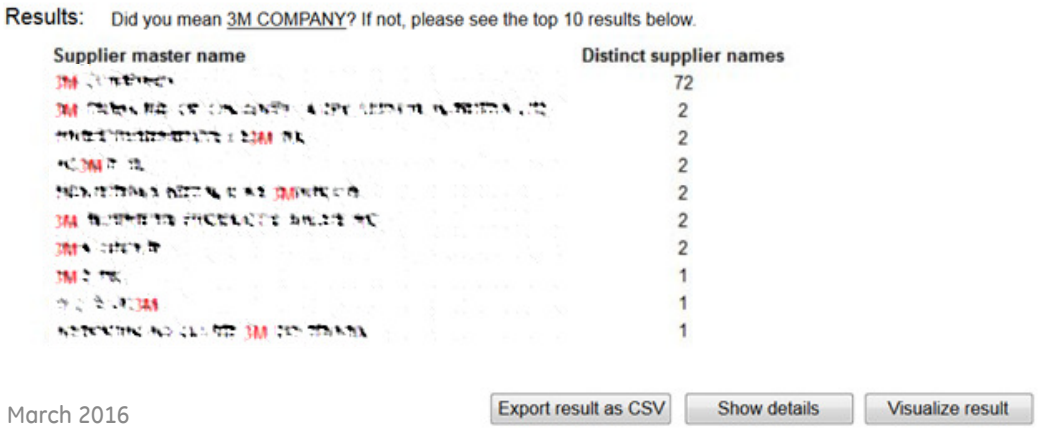
r/s	391,90 x faster...
t/s	710,16 x faster...

```
SELECT *
FROM global_suppliers
WHERE UPPER(supplier_name) LIKE '%GE%ACCESSORY%'
```

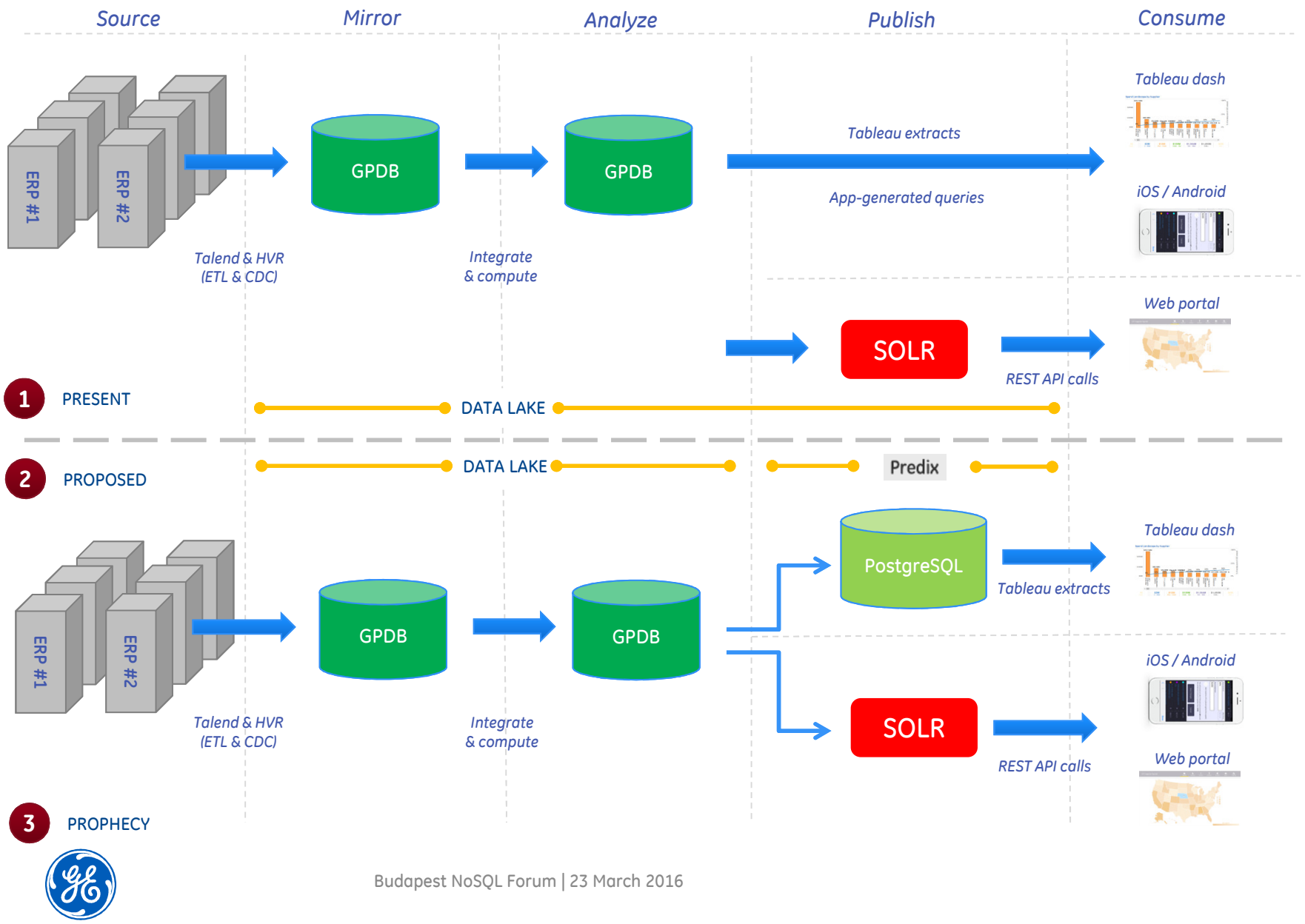
Solr - takeaways

- Solr query configurations are laborious
- Limited statistical, aggregation and analytical capabilities
- Sophisticated search features (e.g. index customizations)
- Rich features (e.g. spelling)
- Powerful UI / admin space
- Facets are cool
- High selectivity data retrieval is FAST





Solr – architecture evolution



Use case highlights

Elasticsearch

Elasticsearch – overview

- Use case
  - Central data repository (aka catalog)
  - Predix / CloudFoundry hosted
  - Geographical queries
  - Diverse queries (but not ad-hoc)
- Tool choice
  - On Predix roadmap
  - Fast prototyping & rewarding development
  - GIS-like capabilities (although not full-featured as PostGIS)
  - Query speed over RDBMS (on high selectivity)

Elasticsearch - more

- Types of data/indexes
  - static data (e.g. hospitals, airports)
    - Scale - n x 100k documents
  - growing data (e.g. weather)
    - Scale – n x 1M documents
- Developer experience
  - Easy to learn & build
  - Output format constrained

Elasticsearch – ELK stack

- Logstash
  - primarily made for log collection
  - sufficient as a ETL tool (with limitations)
  - works well as ES loader
- Kibana
  - visualization tool with web UI
  - customizable via JavaScript
  - limited number of visualization types
  - worth a look if there's no visualization tool in your stack

Tool comparison (recap)

Recap - search tool comparison

Attribute	Solr	Elasticsearch
open source	fully	partially
full text search	yes	yes
aggregation, grouping, analytics	no	yes
built-in front-end tool	yes	no
extensibility (custom features)	high	low
Hadoop integration	yes	yes
managed service	no	yes
commercial support	yes	yes
logo design	nice	hmm
learning & development lifecycle	slower	faster

Recommended goody

Nginx

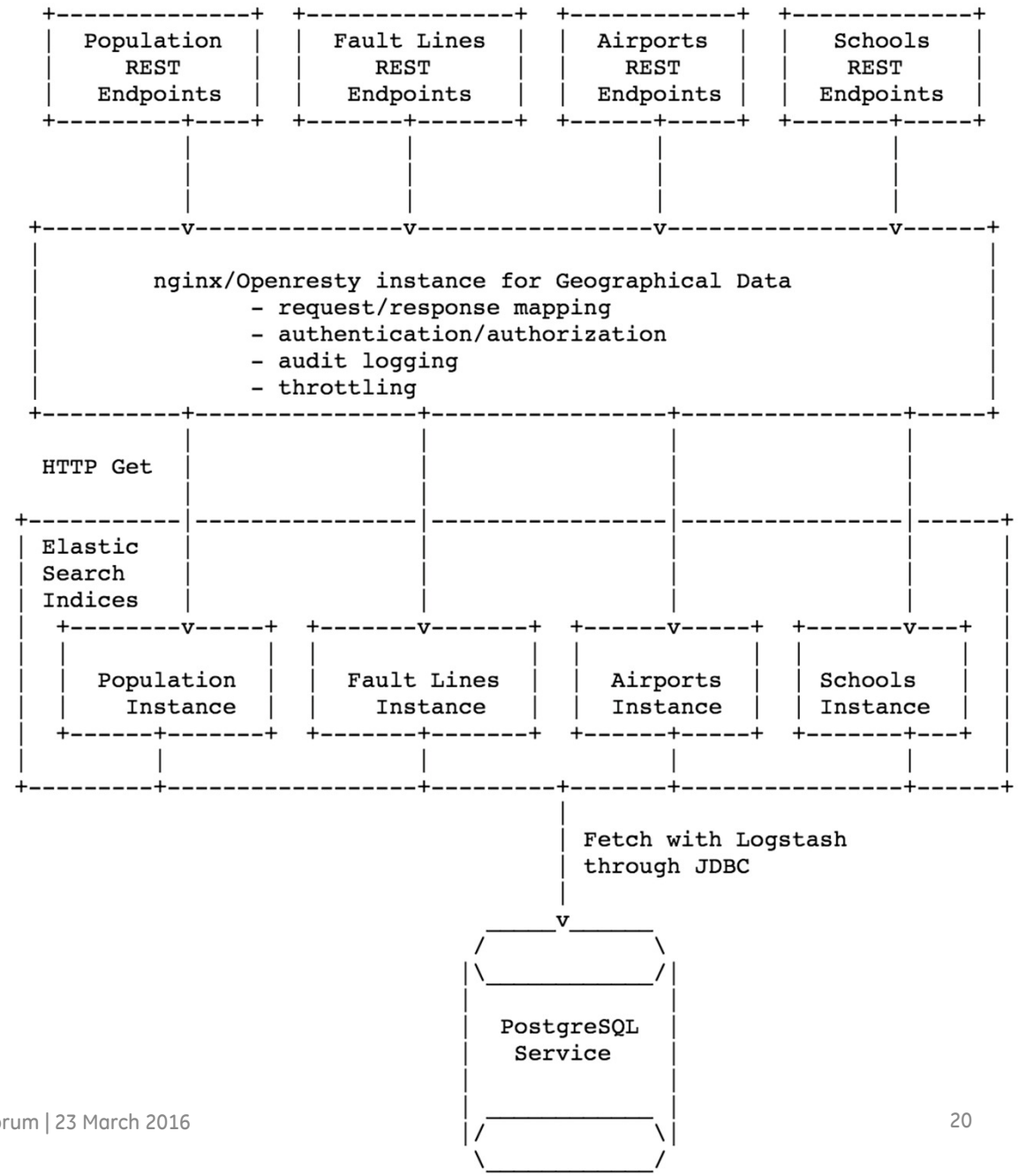
- webserver
- mail/web/reverse proxy (!)
- load balancer
- autenthication & access control
- API version control
- audit & capture API usage

Openresty

Openresty = nginx + Lua + extra libraries

- enables Lua scripting, such as
  - transform Elasticsearch output
  - handle Predix environment variables
- serving data from RDBMS

Asciitecture



Disclaimer / copyrights

General Electric Company reserves the right to make changes in specifications and features, or discontinue the product or service described at any time, without notice or obligation. These materials do not constitute a representation, warranty or documentation regarding the product or service featured. Illustrations are provided for informational purposes, and your configuration may differ.

This information does not constitute legal, financial, coding, or regulatory advice in connection with your use of the product or service. Please consult your professional advisors for any such advice.