
Lucidworks

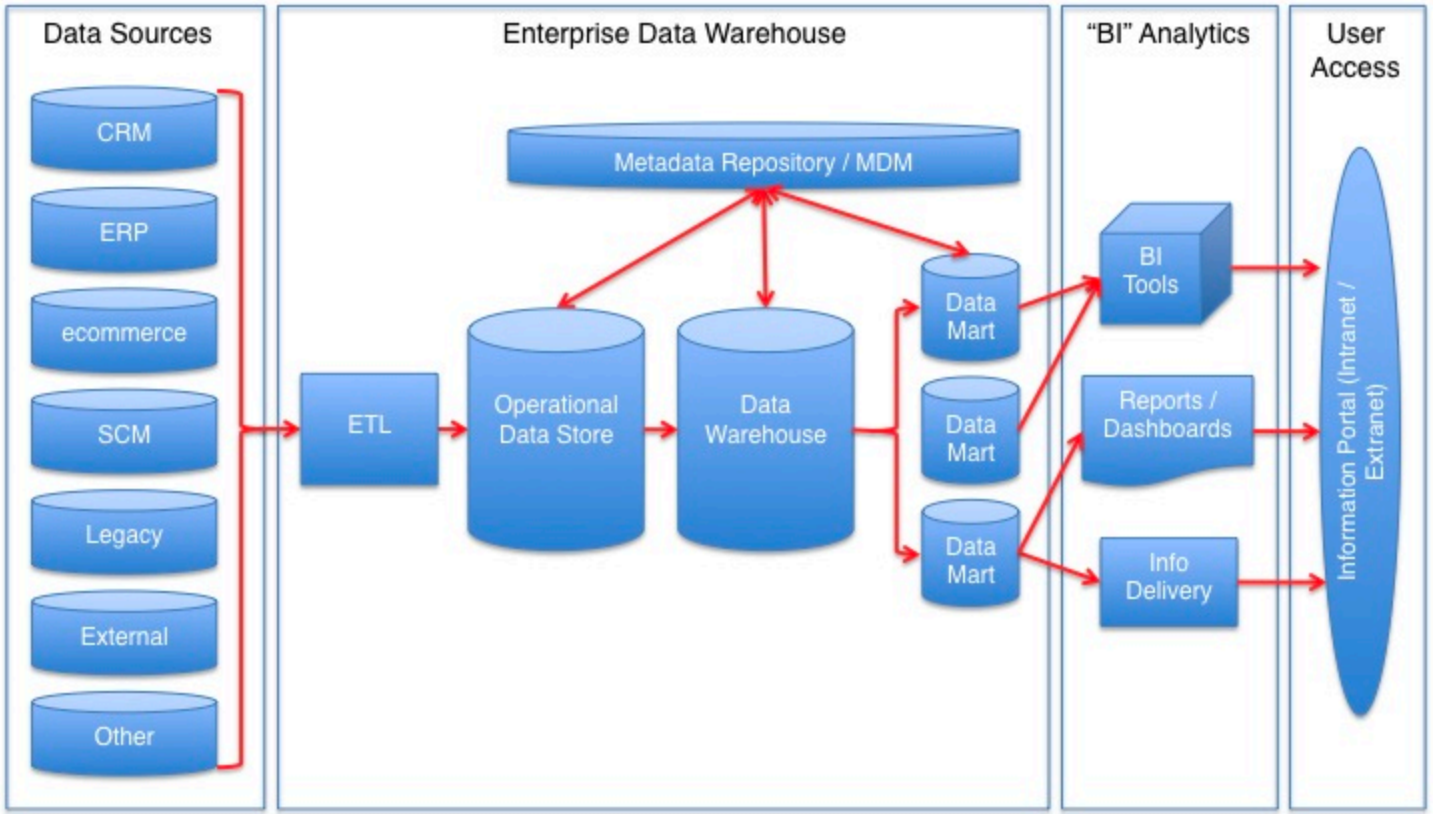
Search as the New Data Warehouse?

Grant Ingersoll
CTO, Co-founder
Lucidworks

FUSION
App Studio

FUSION
Server

FUSION
AI



https://en.wikipedia.org/wiki/Data_warehouse#/media/File:Datawarehouse_reference_architecture.jpg

What is a Data Warehouse?

A data warehouse is a central repository of information that can be analyzed to make better informed decisions. Data flows into a data warehouse from transactional systems, relational databases, and other sources, typically on a regular cadence. Business analysts, data scientists, and decision makers access the data through business intelligence (BI) tools, SQL clients, and other analytics applications.

Data and analytics have become indispensable to businesses to stay competitive. Businesses use reports, dashboards, and analytics tools to extract insights from their data, monitor business performance, and support decision making. These reports, dashboards and analytics tools are powered by data warehouses, which store data efficiently to minimize I/O and deliver query results at blazing speeds to hundreds and thousands of users concurrently.

https://aws.amazon.com/data-warehouse/

Hallmarks of DWs

- Designed to power data marts, BI tools and other analytics with a focus on business enablement, not storage

- Multiple data sources (batch) loaded into DW using ETL tools
  - Separation of concerns with transactional systems
- Highly structured, schema-based design geared around notions of facts, dimensions and aggregations
- Designed for fast querying by hundreds to a few thousands of users

- Often slow to respond to changing business needs

Hallmarks of Data Lakes

- Collect all data types, often in raw form from a variety of data sources
- Multiple data sources dumped into DL using ETL tools
- Schema on read*
- Designed for machine learning, discovery, profiling, analytics

- Often turns into a data swamp since everything gets dumped there and no one knows what's in it

Common Technical Requirements

- Full SQL support with ODBC/JDBC connections
  – Joins, aggregations, grouping, math/stats, UDFs, machine learning
- Schema? Support for many different data types
  – Numeric, Enumerated/Categorical, Text, Spatial, Custom, BLOB
- Data Governance and Metadata Management (MDM)
- Security
- ETL?
- Scalable
- Interfaces with BI and data science tools

The Case Against

- Governance and MDM typically lacking from open search stacks
  - Not particularly hard to add, however
- Schema on read sorta, kinda, maybe
- Traditional RDBMs concepts don't always port
  - Tables, star/snowflake schemas, data modeling
- Lack of training/understanding on how to tune search engines for DW workloads

The Case Against

- Data and workloads are ideally suited for traditional RDBMs (numeric-ish, set-based operations)
  - Dense vectors not quite 1st class yet

- Specific data workloads at scale
- Large scale “S3”-like data storage*
- No Filesystem semantics (data lake scenario)
- Many DBs have added search capabilities, enabling more powerful queries

The Case for Search as DW/DL

- Most major open source (Solr, Elastic) engines now support SQL
  - Fusion SQL supports SQL + Solr query operators
  - Fusion also supports JDBC/ODBC
- Most support joins, grouping

SQL

```
WITH sigs_with_filters AS (
   SELECT c.query as query,
          c.doc_id,
          q.filters_s as filters,
          c.type,
          c.ref_time,
          c.count_i,
          c.timestamp_tdt,
          c.weight_d
     FROM ${inputCollection} c
 LEFT JOIN (SELECT id, filters_s FROM ${inputCollection} WHERE type='response') q ON q.id =
c.fusion_query_id
    WHERE c.type IN (${signalTypes})
 ), signal_type_groups AS (
     SELECT SUM(count_i) AS typed_aggr_count_i,
            query,
            doc_id,
            type,
            filters,
            time_decay(count_i, timestamp_tdt) AS typed_weight_d
       FROM sigs_with_filters
   GROUP BY doc_id, query, filters, type
```

Analytics in a Search DW

- Most have added major in-index streaming and analytical functions beyond faceting and top-*N* retrieval
  - Solr's Streaming Expressions has large library of math functions designed for interactive analysis (relies on Apache Commons Math)
    - see Joel Bernstein's talk from Activate: https://www.youtube.com/watch?v=Seawc8qWYLI
  - Elastic is fairly limited comparatively, but may be sufficient for simple aggregations

Solr's Math Streaming Expressions

https://lucene.apache.org/solr/guide/7_7/streaming-expressions.html

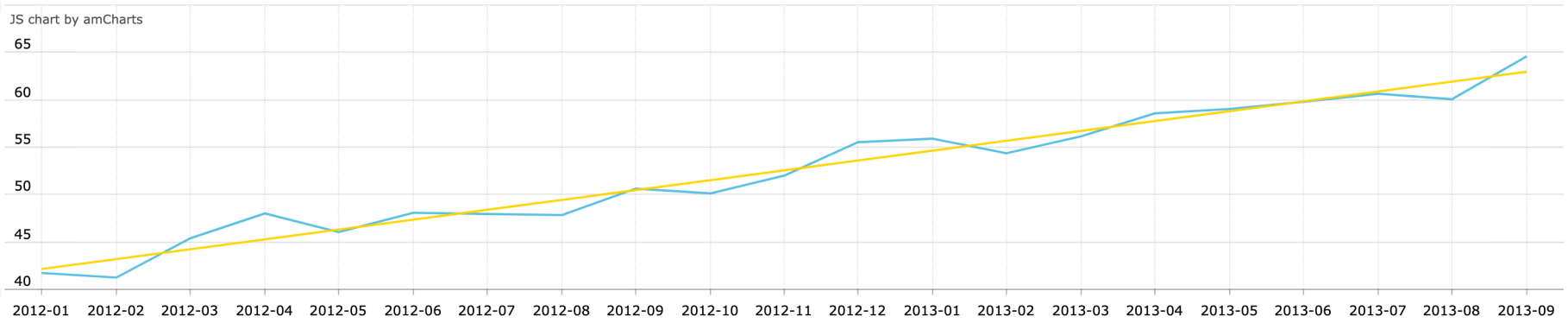
Vector/Matrix Math	Text Analysis
Statistics	Probability
Monte Carlo Simulations	Linear Regression
Curve Fitting	Time Series
Interpolation, Derivatives, Integrals	Digital Signal Processing
Machine Learning	Computational Geometry

```
let(timeline=jdbc(sql="
                    select date_format(tdate_dt, 'yyyy-MM') as month, max(stock_priceB_d) as price
                        from testapp
                    where tdate_dt >= '2012-01-01' and tdate_dt < '2013-10-01'
                        group by month

                                ("),
    x=col(timeline, month),
    y=col(timeline, price),
    n=natural(length(y)),
    r=regress(n, y),
    p=predict(r, n),
    zplot(x=x, y=y, y1=p))
```

FINISHED

settings



Number of results: 21.

Took 0 sec. Last updated by anonymous at March 30 2019, 4:54:37 PM.

Other Factors

- Most have tight integrations with Spark for large-scale data-lake style operations
- Most engines are scalable, fault-tolerant
- Flexible schema and schemaless approaches enable flexibility
- Lucene-based engines have a variety of storage mechanisms available based on data properties and workloads
  – Columnar, dense vs. sparse data
- You have multi-structured data (AKA text and metadata)

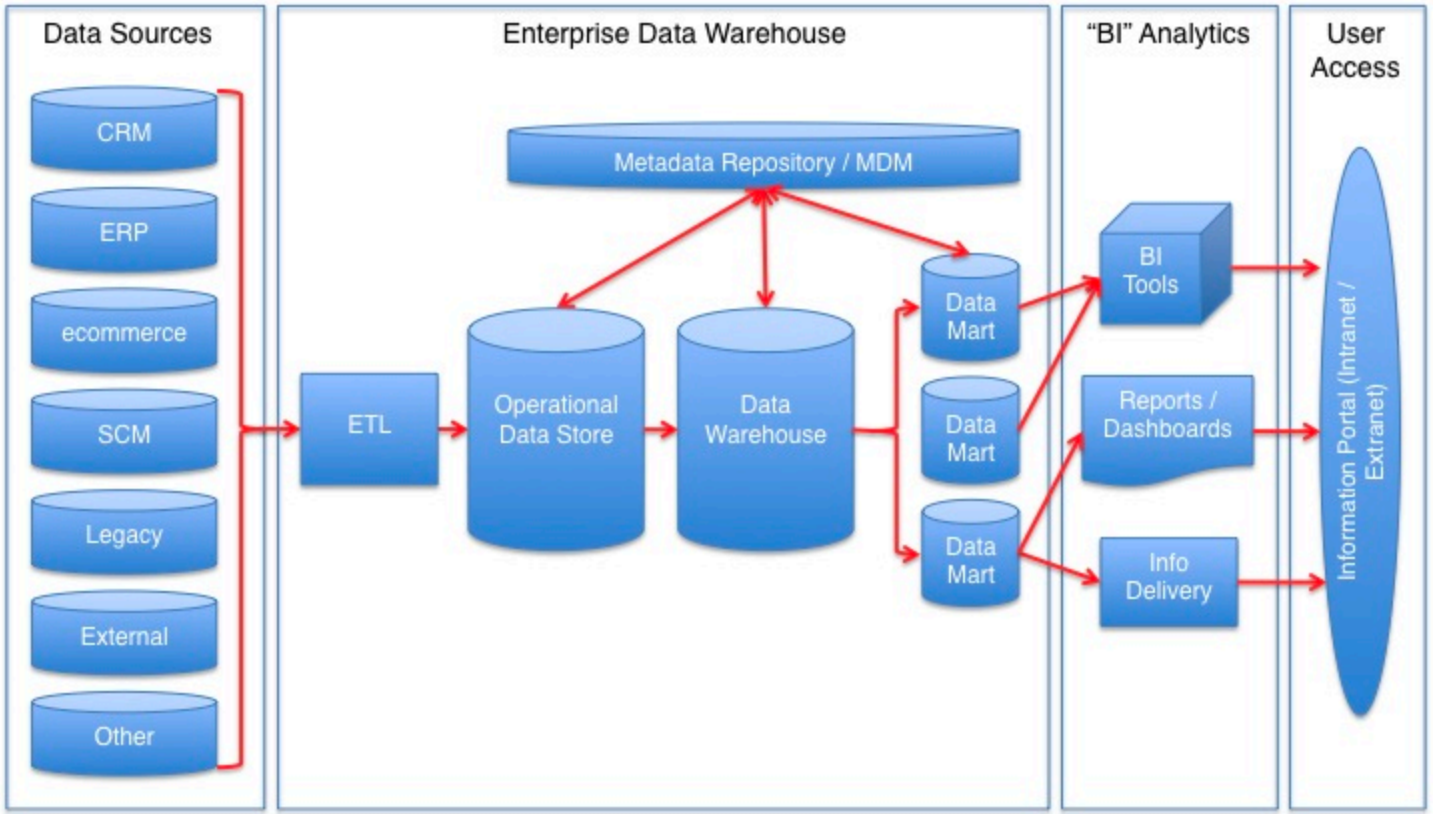
Fusion Server

Highly scalable search engine and NoSQL datastore

- Trillions of data objects - any source, any type
- Thousands of queries per second from thousands of concurrent users
- Full text search, Content Extraction, SQL capabilities
- End-to-end security
- Advanced spatial search and analytics
- Extensible

Solr

APACHE Spark™



https://en.wikipedia.org/wiki/Data_warehouse#/media/File:Datawarehouse_reference_architecture.jpg

Lucidworks

Connectors


Logs


File


Web


Database


Box/ Dropbox


Google Drive


Elasticsearch


Sharepoint


Slack

...


Hadoop


SDK

FUSION AI

NLP: NER, Phrases, POS

Document Classification

Anomaly Detection

Clustering

Topic Detection

FUSION Server

Search Engine & Data Processing

Connectors

ETL Pipelines

SQL Engine

Rules Engine

Scheduling & Alerting

Solr

Apache Spark

FUSION AI

Query Pipelines

Query Intent Detector

Automatic Relevancy

Signals & Query Analytics

Recommenders

A/B Testing

FUSION App Studio


Enterprise Search Template


Analytics Template


Customer 360 Template


Tableau, Jupyter, et. al.

Deploy and Operations


Extensible


Scalable


CDCR


Security Built In

Did I mention you can still search?

select * from tables where … **sort** by

***importance*(*inputs*)**

***importance()* =**

- Proactive
- Relevant/Related
- Timely/Real-time
- Personal/Novel/Diverse
- Free from bias
- Context aware
- Security/Compliance aware
- Learned

lucidworks.com
g@lucidworks.com, @gsingers

We’re hiring!