ta Analytics Reference Architectures – Bi on Facebook, LinkedIn and Twitter

January 18, 2016

nsformationpeople.com/authors/birendra-kumar-sahu/)

Birendra Kumar Sahu
FirstHive
rmationpeople.com/supplier_directory/firsthive/)

tHive

about Birendra Kumar Sahu
rmationpeople.com/authors/birendra-kumar-sahu/)

Big Data is becoming a new technology focus both in science and industry, and motivate technology shift to data centric architecture and operational models.

There is a vital need to define the basic information/semantic models, architecture components and operational models that together comprise a so-called Big Data Ecosystem.

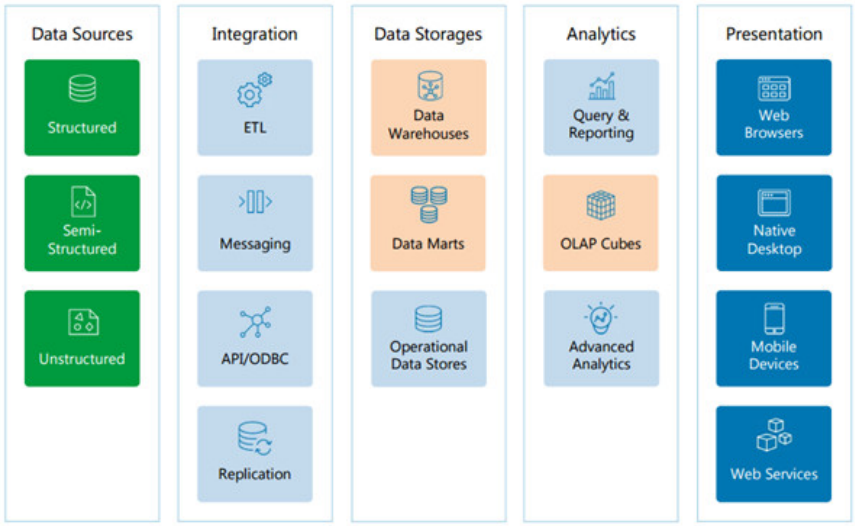
Extended Relational Reference Architecture:

(https://www.thedigitalt

Popular Now

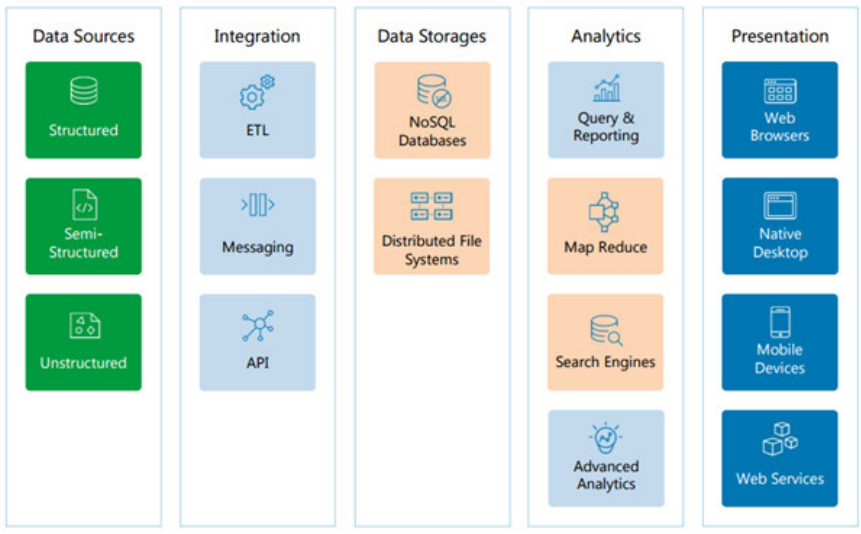
The Case For Digital Transformation
An Executive Summary: Leading Dig

components with pink blocks cannot handle big data challenges.



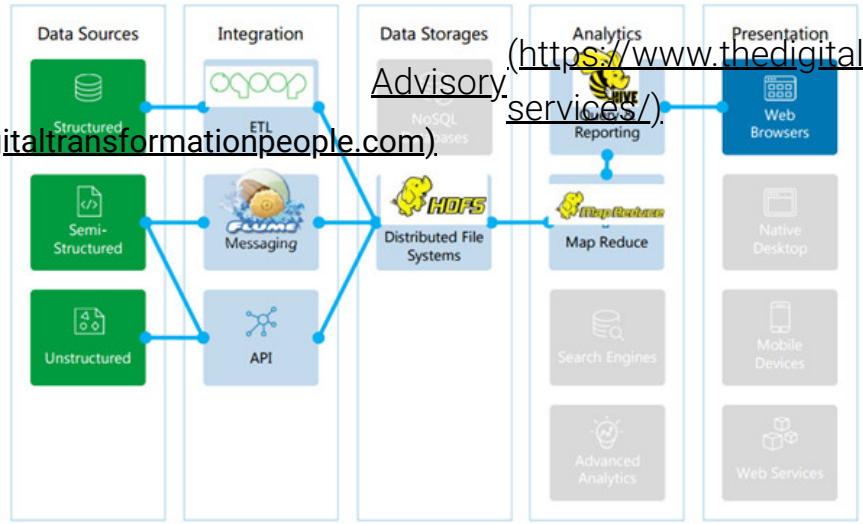
Non-Relational Reference Architecture:

This is more about Non-Relational Reference Architecture but still components with pink blocks cannot handle big data challenges completely.



Data Discovery: Big Data Architecture

This is more about Hadoop based Big Data Architecture which can be handle few core components of big data challenges but not all (like Search Engine etc)



Data analytics Architecture adopted by Facebook:

Data analytics infrastructure at Facebook has been given below. Facebook collects data from two sources. Federated MySQL tier contains user data, and web servers generate event based log data. Data from the web servers is collected to Scribe servers, which are executed in Hadoop clusters.

The Scribe servers aggregate log data, which is written to Hadoop Distributed File System (HDFS). The HDFS data is compressed periodically, and transferred to Production Hive-Hadoop clusters for further processing. The Data from the Federated MySQL is dumped, compressed and transferred into the Production Hive-Hadoop cluster.

models-roadmaps-for-change/)

Delivery
Data Asset Management (DAM) (https://www.thedig asset-management-dam/)

Strategy & Innovation
The Innovation Management Theory Evolution Map (https://www.thedig and-innovation/the-innovation-management-theory-evolution-map/)

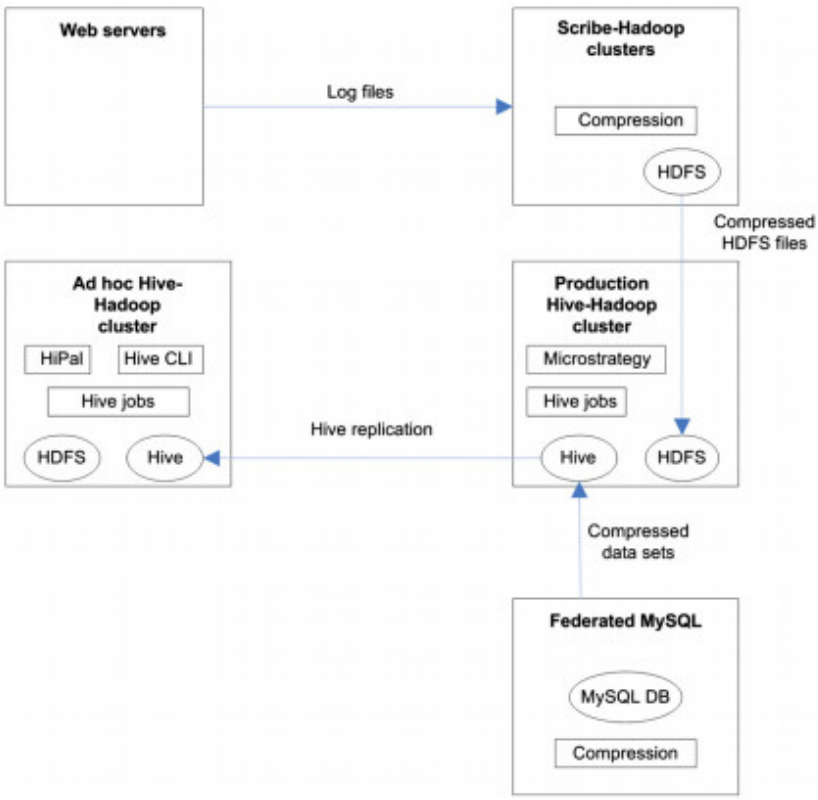
Related Articles

Data & Analytic

Facebook uses two different clusters for data analysis. Jobs with strict deadlines are executed in the Production Hive-Hadoop cluster. Lower priority jobs and ad hoc analysis jobs are executed in Ad hoc Hive-Hadoop cluster. Data is replicated from the Production cluster to the Ad hoc cluster.

The results of data analysis are saved back to Hive-Hadoop cluster or to the MySQL tier for Facebook users. Ad hoc analysis queries are specified with a graphical user interface (HiPal) or with a Hive command-line interface (Hive CLI).

Facebook uses a Python framework for execution (Databee) and scheduling of periodic batch jobs in the Production cluster. Facebook also uses Microstrategy Business Intelligence (BI) tools for dimensional analysis.



Data analytics Architecture adopted by LinkedIn:

The data analytics infrastructure at LinkedIn has been given below. Data is collected from two sources: database snapshots and activity data from users of LinkedIn.

The activity data comprises streaming events, which is collected based on usage of LinkedIn's services. Kafka is a distributed messaging system, which is used for collection of the streaming events.

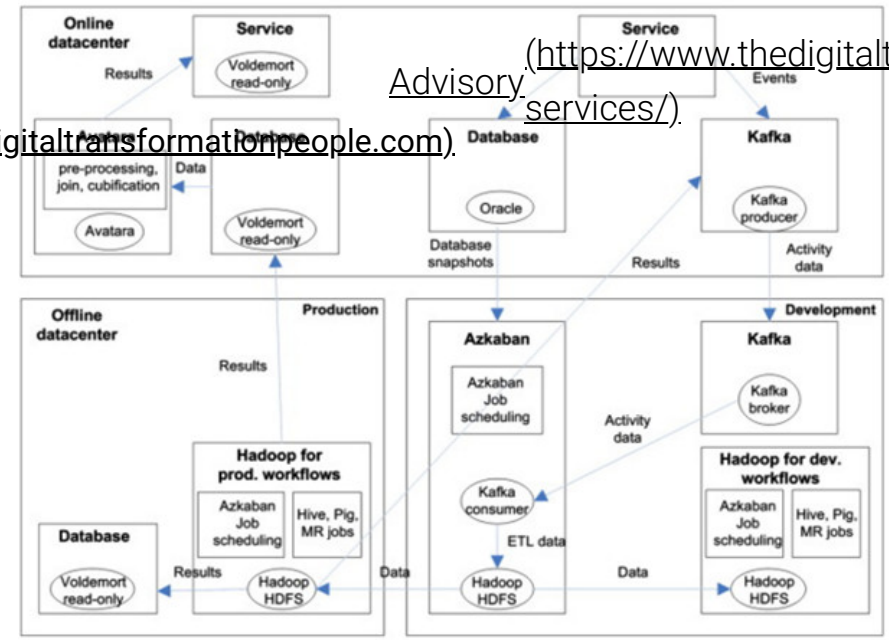
Kafka producers report events to topics at a Kafka broker, and Kafka consumers read data at their own pace. Kafka's event data is transferred to Hadoop ETL cluster for further processing (combining, de-duplication).

Data from the Hadoop ETL cluster is copied into production and development clusters. Azkaban is used as a workload scheduler, which supports a diverse set of jobs.

An instance of Azkaban is executed in each of the Hadoop environments. Scheduled Azkaban workloads are realised as MapReduce, Pig, shell script, or Hive jobs. Typically workloads are experimented in the development cluster, and are transferred to the production cluster after successful review and testing.

Results of the analysis in the production environment are transferred into an offline debugging database or to an online database.

Results may also be fed back to the Kafka cluster. Avatara is used for preparation of OLAP data. Analysed data is read from the Voldemort database, pre-processed, and aggregated/cubificated for OLAP, and saved to another Voldemort read-only database.



Data analytics Architecture adopted by Twitter:

In the Twitter's infrastructure for real-time services, a Blender brokers all requests coming to Twitter. Requests include searching for tweets or user accounts via a QueryHose service. Tweets are input via a FireHose service to an ingestion pipeline for tokenization and annotation.

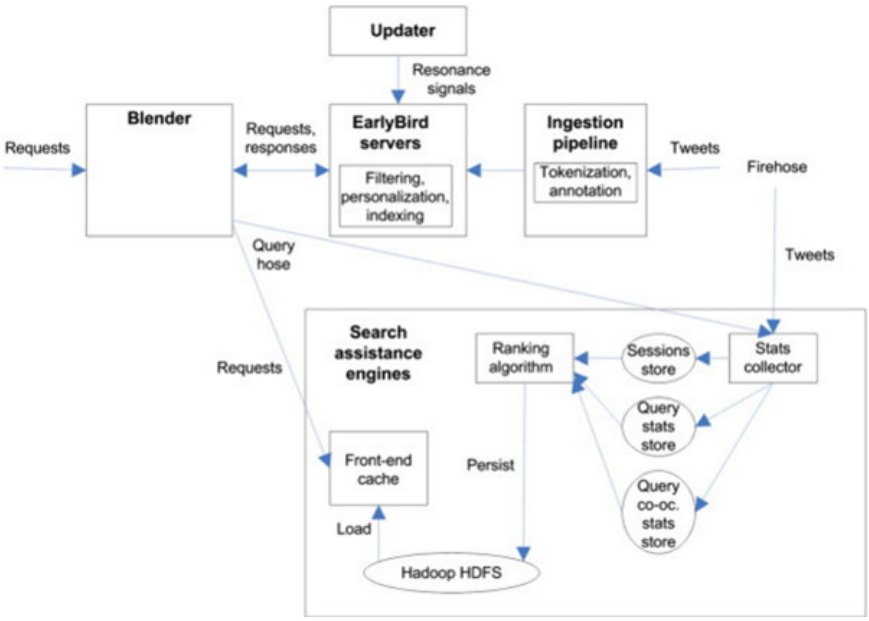
Subsequently, the processed tweets enter to EarlyBird servers for filtering, personalization, and inverted indexing .

The EarlyBird servers also serve incoming requests from the QueryHose/Blender. The EarlyBird is a real-time retrieval engine, which was designed for providing low latency and high throughput for search queries.

Additionally, search assistance engines are deployed. Stats collector in the Search assistance engine saves statistics into three in-memory stores, when a query or tweet is served.

User sessions are saved into Sessions store, statistics about individual queries are saved into Query statistics store, and statistics about pairs of co-occurring queries are saved into Query co-occurrence store.

A ranking algorithm fetches data from the in-memory stores, and analyses the data. The results of analysis are persisted into Hadoop HDFS. Finally, Front-end cache polls results of analysis from the HDFS, and serves users of Twitter.



Twitter has three streaming data sources (Tweets, Updater, queries), from which data is extracted. Tweets and queries are transmitted over REST API in JSON format.

Thus, they can be considered as streaming, semi-structured data. The format of data from Updater is not known (streaming data source). Ingestion pipeline and Blender can be considered as Stream temp data stores.

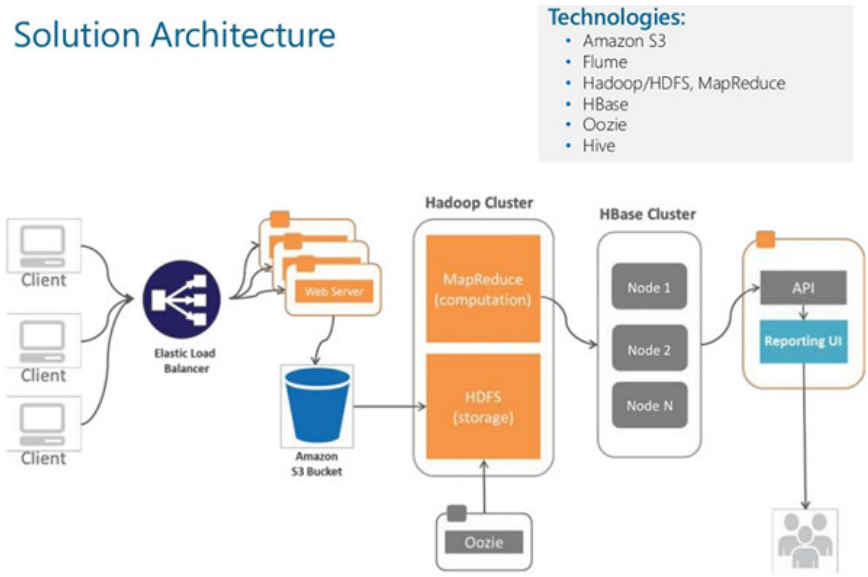
Tokenization, annotation, filtering, and personalization are modelled as stream processing. EarlyBird servers contain processed stream-based data (Stream data store). Stats collector is modelled as stream processing.

The statistical stores may be considered as Stream data stores, which store structured information of processed data. The ranking algorithm performs Stream analysis functionality.

Hadoop HDFS storing the analysis results is modelled as a Stream analysis data store. Front end cache (Serving data store) serves the End user application (Twitter app).

Reference: Reference Architecture and Classification of Technologies by Pekka Pääkkönen and Daniel Pakkala (facebook, twitter and linkedin Reference Architecture mentioned here are derived from this publication)

AWS cloud based Solution Architecture (ClickStream Analysis):



(https://www.thedi

CTO & Vice P

(https://www.thedigitaltransformationpeople.com/authors/birendra-kumar-sahu/)

(https://www.the

Read more articles tagged: Big Data (https://www.thedigitaltransformationpeople.com/channels/tag/big-data/), Featured (https://www.thedigitaltransformationpeople.com/channels/tag/featured/)

Terms of Use (https://www.thedigitaltransformationpeople.com/terms-of-use/)

Medium (https://medium.com/the-digital-transformation-people)

YouTube (https://www.youtube.com/channel/UCPDv4OvDD4Z8uQllQ78PFPg)

GooglePlus (https://plus.google.com/b/100832677067367192859/100832677067367192859/gmbpt=true&pageId=100832677067367192859&hl=en-GB)

(https://www.thedigitaltransformationpeople.com/inter

(https://www.thedigitaltransformationpeople.com

Digital Transformation Consultants/)

services/)

Talent Solutions

Executive Search (https://www.thedigitaltransformationpeople.com/digital-transformation-talent/)

Interim Consultancy (https://www.thedigitaltransformationpeople.com/interim-consultancy/)

Career Opportunities

Candidate Registration (https://www.thedigitaltransformationpeople.com/candidate-registration/)