Comparison of SmartData Fabric® with Cloudera® and Hortonworks®

Revision 2.1

Contents

Introduction

Cloudera® and Hortonworks® are fast becoming the top Hadoop distributions for enterprises, because they are basically free and bundled with the necessary add-ons. They both have active open source communities and commercial support. The main difference between the two is that Cloudera charges for some of the add-ons, whereas, Hortonworks is all open source and free, otherwise, they are almost identical. The alternatives are to either (a) buy similar bundled packages from MapR®, Oracle®, Pivotal®, Teradata® or other vendors, which include support, or (b) build and support own systems using components from open and commercial sources. The core Hadoop components are free and open source, but bundling and support are significant costs of any Hadoop installation, particularly considering the lack, and therefore cost, of Hadoop experience and expertise, and thus the attraction of bundled packages such as Cloudera and Hortonworks.

WhamTech's SmartData Fabric® (SDF) distributed data virtualization, integration, interoperability, analytics and security platform is not Hadoop-based, but could potentially run on Hadoop. WhamTech defines "Big Data" as being all data that is of use to the enterprise, including enterprise systems, Cloud, Hadoop and other Big Data systems, public and commercial sources. Instead of copying or moving all data to a single Hadoop system, WhamTech SDF brings Hadoop-type and more conventional capabilities to data, leaving it in its source. WhamTech SDF uses federated adapters to build and maintain independent, distributed and advanced indexes, usually in near real-time, which are used to process queries and search. The adapters reside transparently in between data sources and middleware/applications. Similar to Hadoop, WhamTech SDF scales through distributed parallel processing, but enables considerably more capabilities. It is based on standards – drivers, SQL and data views such as FHIR APIs, regardless of how or where data is stored in sources. It can be managed in a Cloud infrastructure such as Amazon Web Services™, IBM Bluemix™ or OpenStack™, and other Windows® and Linux™ platforms.

WhamTech SDF is designed to complement and leverage existing IT architectures, systems, data sources, tools and applications – not to replace them, unless some improvement is being sought.

Cloudera, Hortonworks and WhamTech SDF Comparison Table

Features	Cloudera/Hortonworks	WhamTech SDF
DATA-RELATED		
Source	Batch copy - “Data Lake/Reservoir” concept	Remains in data sources, almost any and all, including Hadoop when used a data source
Data quality and standardization addressed upfront	Additional and separate post-Data Lake copy process - “Data refinery” concept	Yes, applied on indexing and, optionally, on results retrieval
Dynamic data masking, tokenization and encryption (WhamTech favors Format-Preserved Encryption (FPE)) – depends on RBAC	No, a separate process, if at all	Yes, applied on indexing and on results retrieval, depends on RBAC
Master data addressed upfront	No, a separate process, if at all	Yes, a seamless integration with other data
Near real-time updates	Maybe, typically batch, but can be streaming	Yes
Changes allowed	No, append only	Yes
Write back to sources	No	Yes
Data tagged with pointers back to source	Not inherent	Yes, inherent
Data discovery and profiling	No	Yes, inherent
Metadata repository	No	Yes, inherent and distributed
Security	Not inherent, but with additional component(s)	Yes, through indexes

Features	Cloudera/Hortonworks	WhamTech SDF
INDEXES-RELATED		
Indexes – structured, advanced unstructured text, fuzzy, master data, categorization, classification, security, extracted entities and others	Not inherent, but with additional component for unstructured, e.g., Solr, and data copied for structured, e.g., Cassandra	Yes, distributed and columnar, and can be segmented
Unstructured data pre-processed, e.g., entity extraction, categorization and advanced text indexing	No	Yes, as indexes built and updated
Structured and unstructured data security classification	No	Yes, through indexes
Indexed views – virtual and material – near real-time updateable – can be hierarchical - for pre-joins, routine queries, triggers, pre-aggregations and pre-calculations	No	Yes
Link Indexes™ upfront for pre-joins, master data management, link analysis, degrees of separation queries, etc.	No	Yes
QUERIES-RELATED		
High performance, parallel distributed query processing	Yes, segmented full table scans	Yes, columnar, data source level and data source segment-level
Queries/joins across data segments – no performance concerns	No	Yes, plus, can use pre-joins in Link Indexes
Combined structured queries and unstructured search	Yes, typically all data is unstructured	Yes, can be in one SQL statement

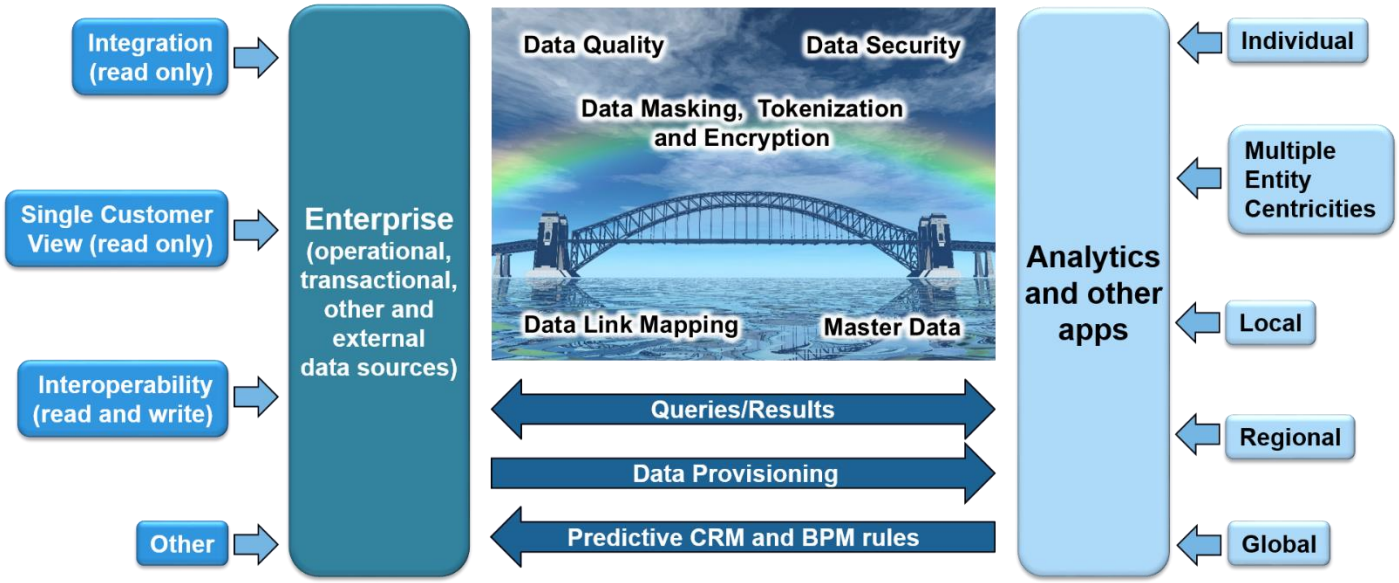
Features	Cloudera/Hortonworks	WhamTech SDF
QUERIES-RELATED (contd.)		
Fully integrated views/result-sets	No	Yes
Triggers/monitoring and event processing	No	Yes, the federated data access partner for Oracle® Event Processing (OEP)
Automatic event correlation, e.g., identifying a cyber-attack through an ontology and establishing standard user behavior	No	Some already with Link Indexes, rest, future
Automatic anomaly detection, e.g., identifying non-standard user behavior and predictive analytics	No	Some already with Link Indexes, rest, future
GENERAL		
Standard drivers	Yes	Yes, JDBC, ODBC, C/C++ and Java clients, Python, Rest API , Web services, data services, pub/sub, RSS and streaming/pipe
Query languages	NoSQL and SQL with additional component	SQL, PLSQL, Native TQL and others through conversion (OQL and SPARQL)
Semantic mapping to a data model (standard or otherwise)	No	Yes, indexes mapped to one or more data models
Interoperability through a standard data model	No	Yes, including writing to data sources

Features	Cloudera/Hortonworks	WhamTech SDF
Link analysis/graph database	No, a separate application, e.g., Titan DB	Yes, optional OEM highly interactive link visualization

WhamTech SDF Complements Hadoop-based Big Data, Cloud, Data/Link Visualization and Other Applications

WhamTech's data virtualization approach provides benefits that can complement a pure Hadoop-based Big Data approach and by doing so, enhances what Hadoop can do and, ultimately, provides significant value over and above Hadoop distributions alone. WhamTech can continuously or on demand provision high quality and additional value-added data, and information to any target, including Hadoop, almost any database, Cloud, data/link analytics/visualization and other applications. The same approach applies to legacy application data transition-migration to these various targets – a combination of inherited WhamTech experience with IBM mainframe VSAM file relational indexing and SQL query processing, and other past projects.

In fact, the approach WhamTech advocates is more along the lines of a Logical Data Warehouse (LDW) or perhaps Logical Big Data (LBD), enabling Cloud-like access to data, regardless of where and how it resides. With a Cloud/data services platform such as Amazon Web Services, IBM Bluemix or OpenStack, WhamTech can enable Hybrid Cloud access to both Cloud/data services-based and on premise data sources. WhamTech bridges the gap between the enterprise and analytics, including analytics running on Big Data systems, as illustrated in the diagram below.



For data provisioning, WhamTech takes care of many processes upfront and thereby enables the following:

- **DATA QUALITY**
  Removes the need to put raw data through a "Data Refinery"/ETL process and then into an analytics database to perform complex analytics, as analytics can be run directly on curated data in Hadoop through SPARK, YARN or similar

- **DATA MASKING, TOKENIZATION AND/OR ENCRYPTION**
  Data can be masked, tokenized and/or encrypted (WhamTech favors Format-Preserved Encryption, i.e., tokenization AND encryption), greatly alleviating security, privacy, regulatory and other concerns

- **DATA RELATIONSHIP MAPPING**
  Can leverage built-in data relationship mapping tags for link analysis and visualization, and other graph database type analytics and/or semantic model representation and queries – no need to compute, just visualize and interact

- **MASTER DATA MANAGEMENT AND TAGGING**
  Leveraging data relationship mapping to match entities within and across data sources, to build and maintain distributed hybrid

master data and associated indexes, using direct matches for simple entities and composite-weighted multi-attribute probabilistic matches for complex entities (whose attributes are simple entities)

- **DATA TAGGING BACK TO SOURCES**
  All data inherently tagged with pointers back to operational/transactional and other enterprise systems and external data sources, Cloud, Web, office documents, email, etc., the gap between analytics and ongoing operations can be bridged, in near real-time (or real-time, depending on performance) – also supports audits

- **DATA SOURCE MONITORING AND EVENT PROCESSING**
  Can set triggers on/monitor data sources through near real-time updates of indexes with event processing that can support alerts, BPM, CRM and other activities at the Hadoop level, e.g., WhamTech is the preferred Oracle® event processing (OEP) partner for federated data access

- **DATA PRE-AGGREGATION, PRE-CALCULATIONS AND PRE-JOINS**
  Can work directly on pre-aggregated, pre-calculated and pre-joined data, instead of much larger quantities of raw data that require these operations in Hadoop or during the Data Refinery process – this is particularly beneficial for data generated from IoT devices, and for BI and analytics

- **AUTOMATIC EVENT CORRELATION**
  YIN – can leverage automatic event correlation, e.g., identifying a cyber-attack through an ontology, establishing standard user behavior and analytics – something that WhamTech has with Link Indexes, but continuing to develop and make automatic

- **AUTOMATIC ANOLMALY DETECTION**
  YANG – can leverage automatic anomaly detection, e.g., identifying non-standard user behavior, predictive analytics and fraud detection

- **IMPROVED BI, REPORTING AND ANALYTICS**
  Population or individual BI, reporting and analytics around curated data, including master data, can dramatically improve results quality and even conclusions

About WhamTech, Inc.

WhamTech, Inc. (WhamTech) is a privately-held US-owned Delaware Corporation established in October 2000 and based in Dallas, Texas. WhamTech's mission is to develop indexed adapter-based distributed virtual data management, master

data management, analytics and security platform software products. WhamTech develops these products to anticipate, meet and exceed the demands of customers seeking an alternative to the conventional approaches of data warehousing, federated data access with conventional adapters, and enterprise search. WhamTech's goal is to provide an improved and more seamless way to work with data, by leaving it in sources and changing the way fundamental and advanced data management is addressed. Most WhamTech adapter products leverage independent, cleansed, transformed and standardized indexes that execute both structured and unstructured queries, and seamlessly and automatically integrate master data management to provide capabilities normally associated with multiple separate solutions, including providing results when data sources are unavailable and for archive.

Information on WhamTech solutions, sales and services, and partnership and investment opportunities can be obtained through whamtech.com.