**CenturyLink Business Objective**

**for Functional Architecture**

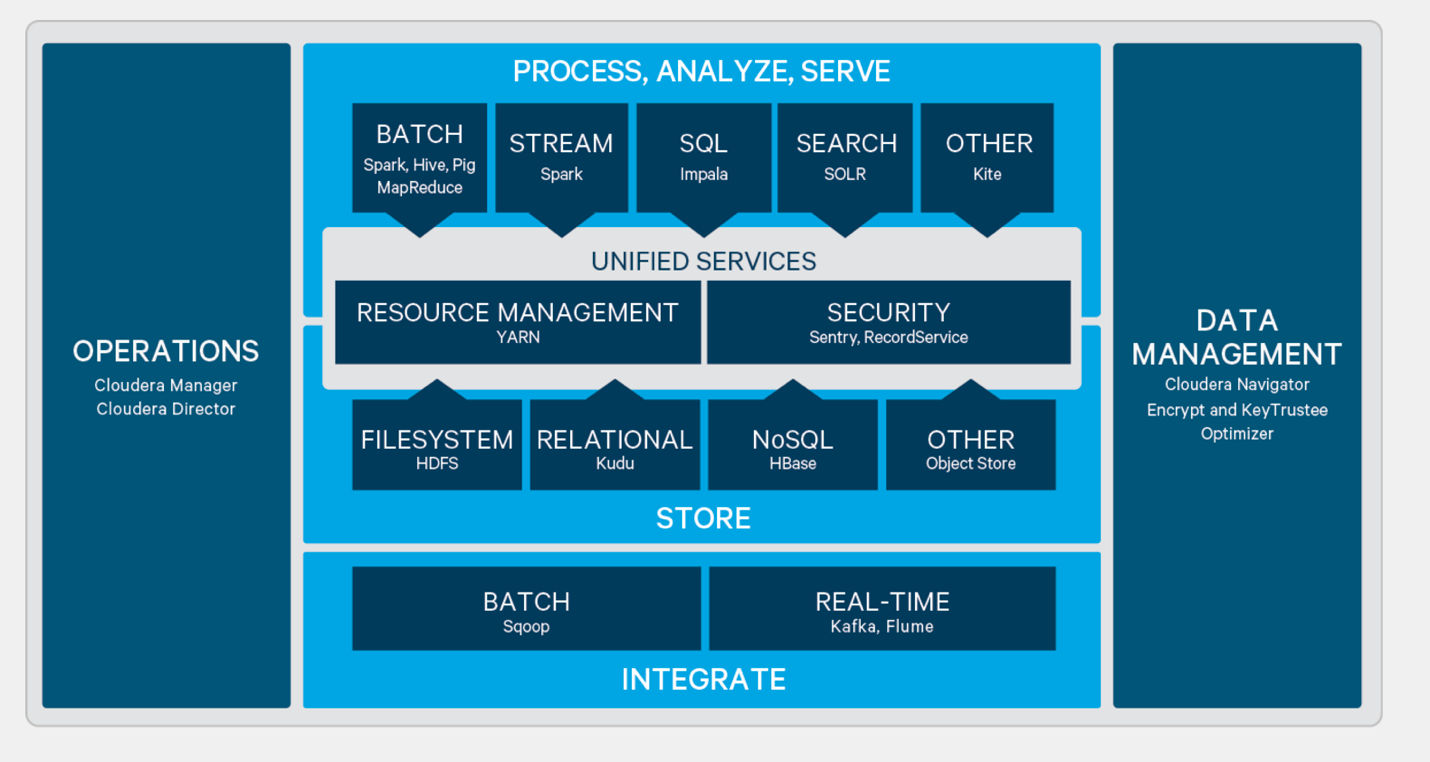
**as of August 28, 2016**

CenturyLink would like to identify several POCs to allow us to bring in providers of big data analytics and test them. Concurrently, CenturyLink wants to test the providers of analytics platforms and related tools.

Many big data software platforms are being sold which do many things. This creates difficulty in framing up the big data platform selection decisions that service a variety of CenturyLink’s business requirements and goals.

For obvious CenturyLink operational efficiencies, business unit collaboration, eliminate duplicate time, money, effort spend, etc., this functional architecture document proposes a set of leading edge tools for CenturyLink big data programs and projects to standardize on. To achieve that, we have to ensure our big data software tool selections meet the business objectives and project success requirements of the wider business for each category of needs.

Some of the first pressures that we will see from our business units will relate to software products like Cask & StreamAnalytix (which business units are interested in), as well as platforms like Splunk & SAP Hana (which business units already use).



**Extract, Transform and Load (ETL) Cloudera certified partner tools**

Informatica

Cloudera Integration Points: HDFS, Hive, MapReduce, ODBC, Pig

Functional Description: Data integration, data quality, data parsing, data lineage, master data management, data archiving and data security capabilities. GUI based ETL. Proprietary product

Syncsort

Cloudera Integration Points: Parcels, HDFS, MapReduce, Spark.

Functional Description: DMX-h: Mainframe and ETL offload to Hadoop. End to end approach to collect, prepare, transform, blend and distribute data. SILQ: Analyzes SQL queries and provides DMX-h recommendations

Pentaho

Cloudera Integration Points: HBase, HDFS, Hive, Impala, JDBC, MapReduce, Pig, Sqoop, YARN.

Functional Description: Data integration and analytics platform. Data integration includes drag and drop ETL, data profiling and data quality. Recently acquired by Hitachi Data Systems. Based off open source

Talend

Cloudera Integration Points: HBase, HDFS, Hive, JDBC, MapReduce, Pig, Spark, Sqoop.

Functional Description: Data integration, quality and management tool. GUI based code generation for drag and drop ETL. Connects different data sources into the Hadoop ecosystem and utilizes the MapReduce framework. Built in connectors to integrate with CDH. Based off open source.

SnapLogic

Cloudera Integration Points: HDFS, YARN.

Functional Description: Delivers a fast, multi-point integration platform that helps enterprises elastically integrate all cloud and on-premise applications, APIS, and disparate data sources for analytics.

**Data Discovery and Analytics, Cloudera certified partner tools**

Zoomdata

Cloudera Integration Points: Hive, Impala, Search, Spark.

Functional Description: Data exploration and self service BI, data visualization platform to allow business users to derive insights data.

Datameer

Cloudera Integration Points: HDFS, Hive, MapReduce, Sentry.

Functional Description: End to end big data analytics platform that combines data integration, analytics and visualization.

Platfora

Cloudera Integration Points: HDFS, Hive, JDBC, MapReduce.

Functional Description: Comprehensive Big Data analytics platform to explore, analyze and visualize Hadoop data. Platfora analyzes data in-memory, pulling data out of Hadoop into a data “lens”.

Tableau

Cloudera Integration Points: Hive, Impala, ODBC.

Functional Description: Data blending, analytics and visualization platform.

Microstrategy

Cloudera Integration Points: Hive, Impala, JDBC, ODBC.

Functional Description: Analytics and visualization platform with the ability to create dashboards and visualization. Can connect to multiple data sources.

Qlik

Cloudera Integration Points: Hive, Impala, ODBC.

Functional Description: Analytics, visualization and dashboard platform with faster/scalable analysis (in-memory) and governing features. Qlikview provides guided analytics, Qliksense provides self-service BI.

SAP

Cloudera Integration Points: Hive, Impala, JDBC.

Functional Description: BI suite that includes data discovery, visualization, reporting, dashboards and applications.

SAS

Cloudera Integration Points: HDFS, Impala, MapReduce, Pig, YARN, JDBC, ODBC.

Functional Description: Data management, exploration, analytics, visualization platform with model deployment and execution.

Trifacta

Cloudera Integration Points: HDFS, Pig.

Functional Description: Data discovery, assessment, shaping, and enrichment platform to prepare data for analytical workloads.

**Operational Analytics, Cloudera certified partner tools**

SAS

Cloudera Integration Points: HDFS, Hive, Impala.

Functional Description: Data mining, statistical analysis, forecasting, text analytics, optimization & simulation platforms for advanced operational analytics.

Datameer

Cloudera Integration Points: HDFS, Hive, MapReduce, Sentry.

Functional Description: End to end big data analytics platform that combines data integration, analytics and visualization.

Cask

Cloudera Integration Points: Cloudera Manager, Parcels, HBase, Hive, Impala, YARN, Zookeeper, JDBC.

Functional Description: Abstraction layer that enables developers to build real-time and batch data applications. CDAP supplies Cloudera Manager Service Descriptions and is packaged as a Cloudera parcel.

RapidMiner

Cloudera Integration Points: HDFS, Hive, Impala, JDBC, Mahout, MapReduce, Parquet, Pig, Spark.

Functional Description: Code free graphical environment for building and deploying predictive analytics in Hadoop. Supports the whole lifecycle with ETL, data preparation, exploration, machine learning and scoring models.

Skytree

Cloudera Integration Points: HDFS, MapReduce, Spark, YARN, Zookeeper.

Functional Description: Scalable machine learning platform that provides data preparation, data modeling & machine learning algorithms.

**EDH for Cybersecurity, Cloudera certified partner tools**

Intel

Functional Description: Duxbury Bay solution focused on network data.

SAS

Cloudera Integration Points: HDFS

Functional Description: Outlier detection. Still in Beta.

Securonix

Cloudera Integration Points: Flume; Hadoop; HBase; Hive; Hue; Impala; Parquet; Search; Spark; YARN; Zookeeper.

Functional Description: Purpose built security intelligence platform that plugs in and transforms all areas of security management through advanced analytics and actionable intelligence. Provides signature-less Behavior Based Threat Detection technology and risk based access outlier detection technology and fuzzy logic based identity correlation engine.

Caspida

Cloudera Integration Points: HBase, HDFS, MapReduce, Spark, YARN.

Functional Description: Threat detection and prevention capabilities using advanced user, application, device and data-aware behavior models powered by machine learning algorithms.

Fortscale

Cloudera Integration Points: Cloudera Manager CSD, HDFS, Impala, Kite

Functional Description: Machine learning big data analytics to detect and mitigate user based security threats. Focused on making the security analysts' life easier.

E8 Security

Cloudera Integration Points: Sentry, Flume, HBase, HDFS, Impala, MapReduce, Pig, Search, Spark, YARN, Zookeeper.

Functional Description: Machine learning for threat detection. Compliments SIEM, IDS etc. Includes visualization.

Niara

Cloudera Integration Points: Cloudera Manager, Flume, HBase, HDFS, Kite, MapReduce, Parquet, Spark, YARN.

Functional Description: Niara is building a security monitoring platform that helps discover compromised users and malicious insiders, efficiently prioritizes alerts for rapid investigation and supports advanced threat hunting efforts.

Splunk

Cloudera Integration Points: HDFS, MapReduce.

Functional Description: Platform for machine data. Anomaly detection relies on pre-configured correlation searches. Correlation searches can reference threat intelligence information via the Splunk App for Enterprise Security or via direct import from an intelligence source.

ThreatStream

Cloudera Integration Points: HDFS, Impala

Functional Description: Security intelligence is aggregated from multiple sources, de-duped, ranked, normalized and then fed as a clean stream to the SIEM.

Platfora

Cloudera Integration Points: HDFS, Hive, JDBC, MapReduce.

Functional Description: Comprehensive Big Data analytics platform to explore, analyze and visualize Hadoop data. Platfora analyzes data in-memory, pulling data out of Hadoop into a data “lens”.

**IT Operations Analytics, Cloudera certified partner tools**

SAP

Cloudera Integration Points: HDFS.

Functional Description: IT Operations Analytics. Embeds SAP HANA.

Rocana

Cloudera Integration Points: Avro; Cloudera Parcels; Hadoop; HDFS; Hive; Hue; Impala; Kite; Parquet; Solr Search.

Functional Description: IT Operations Analytics. Leverages large portions of EDH stack. Founded by Ex-Clouderans.

**DataFlow Into Hadoop (device level)**

         StreamSets versus NiFi

StreamSets holds architecture-driven advantages over NiFi in a number of areas including simplicity, reliability, flexibility and transparency.  StreamSets is also distro-agnostic and integrates with existing cluster environments.

Simplicity:  you can drag-and-drop ingest pipelines with StreamSets. NiFi requires substantial coding to glue together stages.

Reliability: StreamSets pipelines require minimal schema description. NiFi pipelines are highly sensitive to schema changes and origindata is destroyed on use, making recovery challenging.

Flexibility: Since each StreamSets stage is independent, adapting pipelines is dead easy.  Changes to NiFi often require recoding since processors require explicit schema and structure knowledge.

Transparency: Since StreamSets uses a standard record format, all data can be introspected for deep real-time monitoring.  NiFi fileshare opaque, allowing for only 5 minute averages of certain metrics.

https://www.youtube.com/watch?v=idqCs-CXYaU

**Log Analytics:**

**Splunk is a certified Cloudera partner**[PARTNER WEBSITE](http://www.splunk.com/)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Cloudera Versions** | **Partner Product Name** | **Partner Product Version** | **Interface Components** | **Supports Kerberos** | **Supports Apache Sentry** |
| CDH 5.0 | Hunk 6.0.2 | 6.0.2 | HDFS, MapReduce, Hadoop Streaming | No | Not applicable |
| CDH3u4, CDH3u5 | Connect 1.0 | 1.0 | HDFS, MapReduce |  |  |
| CDH3u4, CDH4.2 | Hunk 6.0 | 6.0 | Hadoop, HDFS, MapReduce | Yes |  |

**Splunk Functions:**

         Searching & Reporting:   Search Head

         Indexing & Search Services:  Indexer

         Local & Distributed Manageement:  Deployment Server

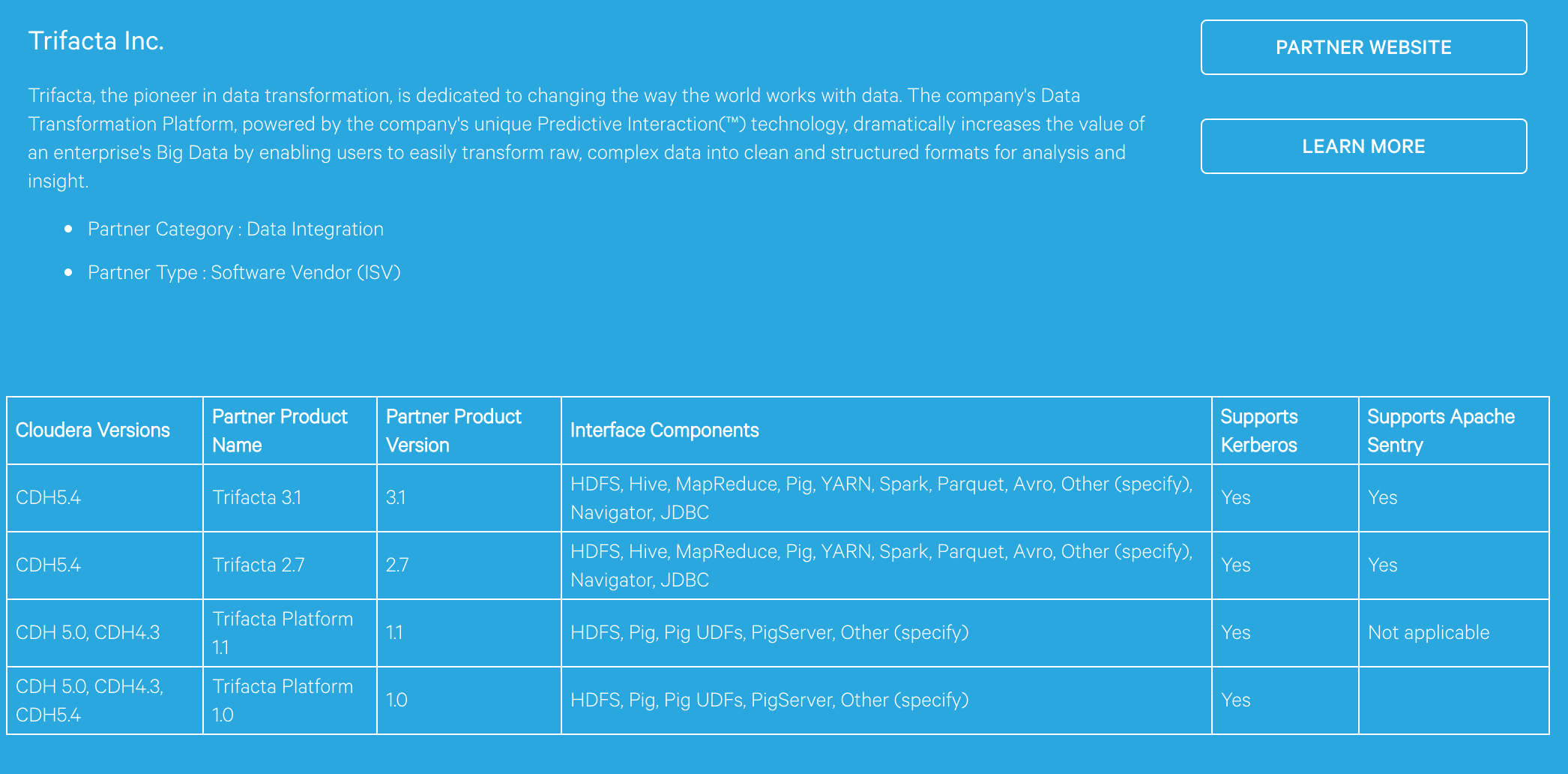
         Data Collection & Forwarding:  Forwarder

         Impetus Analytix

         Data Wrangling

Trifacta

<https://www.youtube.com/watch?v=Ht7oN9wYZ-0&feature=youtu.be&list=PLe-h9HrA9qfA8SSE8uyodaSvsz-gy1ELF>



         Computational Engines

Included with Cloudera distribution:

Spark

**Parallel Complex Event Processing**

<http://softwareengineeringdaily.com/2016/02/04/stream-processing-vs-complex-event-processing/>

**Network/Operations Analytics Platform**

         Cask—application development toward development within Cloudera environment

<https://www.youtube.com/watch?v=LthJ8K31Szc&feature=youtu.be&list=PLe-h9HrA9qfA8SSE8uyodaSvsz-gy1ELF>

<http://go.cloudera.com/Cask_Software_Demo_Reg>

