CenturyLink Technology Solutions Service Guide

Big Data Foundation Services Cloudera

This CenturyLink Service Guide ("SG") sets forth a description of the Big Data Foundation Services - Cloudera ("Service") offerings by CenturyLink, including technical details and additional requirements, if any. This SG is subject to and incorporated into the Service Agreement and Service Schedule between the parties. The specific details of the Service ordered by Customer will be set forth on the relevant Service Order. For avoidance of doubt, any references in the Agreement, Schedules, or Service Orders to SSG, shall mean SG.

Version	Previous	Section Modified	Date
HOS-20141010-SG-BigData-Cloudera	HOS-External-SSG-US-Cloudera Services	New	October 10, 2014

Table of Contents

Service Description

1. **Standard Service Description**: Big Data Foundation Services: Cloudera ("Cloudera" or the "Service") is a Managed Hosting product. At the core of the Service is CDH (the "Software"). Developed by Cloudera ("Provider"), CDH is a distribution of Apache Hadoop and data management tools used by Customers to integrate and analyze information. The standard Service consists of the licensing, installation, configuration, administration, monitoring, maintenance and support for the Software components listed in Section 1.1. The Service Level Agreement (SLA) associated with this Service Guide is the "Managed Hosting Services SLA."

   1.1. **Software Components**: The Software is available in the supported versions with the supported features installed as part of the Service.
      1.1.1. **Supported Versions**:
         1.1.1.1. Hadoop: Cloudera – CDH5
      1.1.2. **Supported Features** (see Definitions)
         1.1.2.1. Sqoop
         1.1.2.2. Pig
         1.1.2.3. Hive
         1.1.2.4. Flume
         1.1.2.5. Oozie
         1.1.2.6. ZooKeeper
         1.1.2.7. HBase – with the purchase of the Basic + HBase package

   1.2. **Licensing**:
      1.2.1. **CenturyLink Provided**: As part of the standard Service, CenturyLink obtains the software licenses required for the Software Components in Section 1.1. All users of the Service are subject to the terms and conditions of the license agreements. Customer is responsible for maintaining up-to-data Cloudera support agreements at all times during the Service Term of its Agreement with CenturyLink. If Customer fails to maintain the support agreement, CenturyLink may, at its option, delay the deployment date or terminate the agreement.
      1.2.2. **Customer Provided**: Customer has the option to obtain software licenses for the Software Components in Section 1.1 separately. If Customer chooses to obtain the licenses, CenturyLink will install them as part of the Service. Customer represents and warrants that a written license agreement exists with Cloudera. Customer agrees to provide CenturyLink with evidence of licensing as requested prior to the deployment date and periodically as requested to update the status of the license. If Customer fails to provide the license agreement, CenturyLink may, at its option, delay the deployment date or terminate the agreement. Customer agrees to add CenturyLink to their support agreement documentation so that Cloudera recognizes CenturyLink as the on-site provider.
      1.2.3. **Indemnification**: Except as otherwise provided by mandatory applicable law without possibility of contractual waiver and notwithstanding anything in the Master Service Agreement or elsewhere, CenturyLink and its affiliated companies shall not be liable to you or indemnify you for any claims of infringement of patent, copyright or other intellectual property right related to the components in Section 1.1.

   1.3. **Installation**: CenturyLink will provide installation tasks marked with an "X" in the CenturyLink column in Table 1.0 Roles and Responsibilities for the supported operating systems and hosting CenturyLink platforms listed below. Customer chooses the operating system and platform.
      1.3.1. **Supported Operating Systems**:
         1.3.1.1. Red Hat Enterprise Linux AS6 version 6.x (64 bit)
      1.3.2. **Supported Hosting Platforms**:
         1.3.2.1. Intelligent Hosting

1.4. **Configuration**: CenturyLink will provide configuration tasks marked with an "X" in the CenturyLink column in Table 1.0 Roles and Responsibilities. Customer chooses the configuration type from the items below:

    1.4.1. **NameNode and Secondary NameNode**: The default implementation of Hadoop deploys a NameNode and a Secondary NameNode to provide Hadoop in a High Availability (HA) mode. If the NameNode fails the Secondary NameNode will take over without a loss of system performance.

    1.4.2. **Secondary NameNode**: Customer can choose the Secondary NameNode configuration. In this configuration, a redundant copy of the NameNode will be maintained. In this configuration, the NameNode is a single point of failure (SPOF). If the process or server becomes unavailable, the cluster as a whole will be unavailable until the NameNode is repaired. Any maintenance on the NameNode server or recovery of the cluster upon hardware failure to another server will have downtime implications to the cluster. The SLA does not apply in this traditional Secondary NameNode configuration for issues related to NameNode.

1.5. **Administration**: CenturyLink will provide administration tasks marked with an "X" in the CenturyLink column in Table 1.0 Roles and Responsibilities.

    1.5.1. **Production Environment**: CenturyLink maintains full root or administrator access on the managed server. Root logins are limited to console access only and logged. Full access to administrative functions by Customer is expressly prohibited in production and staging level environments.

    1.5.2. **Non-production level environments**: CenturyLink understands that in certain cases, Customers need administration functions in order to effectively manage certain applications that are running on the server. In these cases CenturyLink will allow the Customer the necessary access; provided, however, such Customer access will be provided pursuant to separate terms defined by CenturyLink permitting such access

1.6. **Monitoring:** CenturyLink will enable application level monitoring for each environment as they are created within the product configuration.

    1.6.1. **Notification:** CenturyLink will retain primary notification and resolution responsibilities for all automated environment alerts in the Production Environment.

    1.6.2. **Monitors for the Service:** Monitors for the Service can be found in Table 2.0.

1.7. **Maintenance and Support:** CenturyLink will provide maintenance and support tasks marked with an "X" in the CenturyLink column in Table 1.0 Roles and Responsibilities.

    1.7.1. **Change Management:** All changes to the CenturyLink managed applications; systems, network and facilities are subject to CenturyLink's change management process. This process is intended to confirm that work is reviewed for completeness (risk assessment, completed test procedure, metrics for measuring progress, back out procedure, etc.) and accuracy prior to scheduling and implementation.

    1.7.2. **Maintenance Windows:** All times listed under the Scheduled Maintenance Windows are local times and subject to change. CenturyLink will use commercially reasonable efforts to perform routine maintenance only during the Saturday or Sunday defined maintenance windows. See Definitions for additional information.

    1.7.3. **Support**: Support for the Service is provided through the project manager during installation, content migration and Customer validation. At the point of go-live the Service is passed from project management to CenturyLink Service Center for full 24x7 monitoring and management. The point of go-live is when Customer notifies CenturyLink project manager that the environment is ready to go-live.

    1.7.4. **Patch Releases**: CenturyLink compiles, certifies, packages, approves and delivers Software patch versions for installation in a Customer environment. Customer must approve the installation by contacting CenturyLink support. Any Outages directly caused by Customer's failure to accept the implementation of a patch will not be subject to SLA Service Credits.

2. **Customer Responsibilities**: Customer is responsible for all tasks marked with an "X" in the Customer column in Table 1.0 Roles and Responsibilities. Customer acknowledges and agrees that its failure to perform its obligations set forth in Table 1.0 may result in CenturyLink's inability to perform the Services and CenturyLink shall not be liable for any failure to perform in the event of Customer's failure.

2.1. **Provide Contact**: Designate and maintain a Customer Contact during the Service Term (including current contact information). "Customer Contact" means a technical point of contact available 24x7 with sufficient knowledge, authority and access to address configuration issues, event notifications, system or infrastructure modifications and authentication of applicable CenturyLink systems.

2.2. **Prerequisite Services**: The Service is only available to Customers that also purchase the CenturyLink services outlined below, under separate terms and charges.

    2.2.1. **Compute**: Intelligent Hosting (minimum of seven nodes needed to comply with the SLA requirements). CenturyLink has the following configurations designed to support the Service. Customer can choose from the recommendations below or purchase a separate configuration that meets the minimum node requirements.

        2.2.1.1. **Standard Option**:
            HP DL380p Gen8 12-Large Form Factor CTO Server
            8 x HP 8GB 1Rx4 PC3-12800R-11 Kit (64GB RAM)
            12 x HP 1TB 6G SATA 7.2k 3.5" SC MDL HDD (12 TB storage)

        2.2.1.2. **Large Option**:
            HP DL380p Gen8 12-Large Form Factor CTO Server
            8 x HP 8GB 1Rx4 PC3-12800R-11 Kit
            8 x HP 4GB 1Rx4 PC3-12800R-11 Kit (96GB RAM total)
            12 x HP 2TB 6G SATA 7.2k 3.5" SC MDL HDD (24 TB storage)

        2.2.1.3. **High Storage Option**:
            HP DL380p Gen8 12-Large Form Factor CTO Server
            8 x HP 8GB 1Rx4 PC3-12800R-11 Kit (64GB RAM)
            12 x HP 4TB 6G SATA 7.2k 3.5" SC MDL HDD (48 TB storage)

        2.2.1.4. **Performance Option**:
            HP DL380p Gen8 12-Large Form Factor CTO Server
            8 x HP 16GB 2Rx4 PC3-12800R-11 Kit (128GB RAM)
            12 x HP 1TB 6G SATA 7.2k 3.5" SC MDL HDD (12 TB storage)

3. **Additional Services:** At Customer's option and expense Customer can choose to have CenturyLink complete one or more of the tasks in Table 1.0 with an "X" in the Customer column and/or the services listed below. The items can be added to the standard Service (described in Section 1.0) for an additional fee described in a separate Statement of Work ("SOW") or Service Order. Contact a sales representative for more information.

3.1. **Enhanced Features:** The following features are part of the Apache Hadoop ecosystem and require the purchase of additional licenses and support (see Definitions).

    3.1.1. Impala
    3.1.2. Navigator
    3.1.3. Enterprise Backup Disaster Recovery (BDR)

3.2. **Minor Releases**: Minor Releases (see Definitions) require a new installation of the software binaries, which is not included in this Service. CenturyLink recommends that Customer purchase a new managed server (via a Service Order) and migrate data (defined in a SOW) as a best practice.

3.3. **Major Releases**: Major Releases (see Definitions) require a new installation of the software binaries, which is not included in this Service. CenturyLink recommends that Customer purchase a new managed server (via a Service Order) and migrate data (defined in a SOW) as a best practice.

3.4. **HBase Setup and Configuration**

3.5. **HBase Monitoring and Analysis**

Tables and Appendices

Table 1.0 Roles and Responsibilities

Activity	Task	CenturyLink Basic Package	CenturyLink Basic + HBase Package	Customer
Licensing	For CenturyLink provided licenses, purchase, report and document the respective software licenses from the respective big data software vendor	X	X	
	For Customer provided licenses, provide CenturyLink with evidence of licensing as requested prior to the deployment date and periodically as requested to update the status of the license			X
Installation	Perform hardware configuration and validation of server power and processor settings, physical disk layout and RAID sets, network layout and configuration specific to Hadoop	X	X	
	Perform required system level preparations to the Operating System as required for the tuning, and configuration settings for the installation of Hadoop	X	X	
	Review the current/proposed environment for both the expected capabilities and efficiencies in support of the service being defined within the context of this document			X
	Perform product installation as defined by any applicable CenturyLink standard, and as defined by the big data vendor, for the Hadoop cluster	X	X	
	Set up HBase and configure to run in the CDH cluster		X	
	Acquire and install the respective software licenses in the big data product from the respective big data software vendors (as is required for functional operation of the Hadoop cluster)	X	X	
Configuration	Create and manage the respective user accounts in the Hadoop cluster, as is required for customer access, as defined by the applicable CenturyLink standards for managed application software.	X	X	
	Configure Sqoop to connect to relational databases and/or setup or perform data transfers			X
	Write, Author, Review, Implement or Run Pig			X

Activity	Task	CenturyLink Basic Package	CenturyLink Basic + HBase Package	Customer
	MapReduce programs			
	Configure log data collecting, aggregating or movement into HDFS (see Definitions)			X
	Perform data summarization, query or analysis within HBase and/or setup Regions, Replication, Backups, Snapshots, Compression, DataStores, Sharding, StoreFiles, Catalog Tables and/or Data Loading and Write, Author, Review Implement or Run HBase Shell or IRB commands		X	
	Perform custom setup and configuration for the JobTracker(s) node(s),			X
	Perform 3rd party software setup and configuration			X
	Perform the Migration from Hadoop Cluster/Ecosystem Native Internal/Embedded Databases to an External CenturyLink Managed Customer MySQL server			X
	Configure the Hadoop Cluster/Ecosystem to use a CenturyLink Managed Customer MySQL server instead or Native Internal/Embedded Databases	X	X	
	Configure the Hadoop Cluster/Ecosystem to use Kerberos or LDAP for Authentication and/or authorization of users			X
	Perform its own validation, User Acceptance Testing (UAT), Site Acceptance Testing (SAT) of the cluster and applications that comprise its database server as a whole prior to going live, provide confirmation that system is ready to go live			X
	Configure monitoring	X	X	
Administration	Review Data Node performance and perform appropriate OS and Application level tuning per CenturyLink and the Hadoop vendors standards, upon request of Customer	X	X	
	Review Control Node performance and perform appropriate OS and Application level tuning per CenturyLink and the Hadoop vendors standards, upon request of Customer	X	X	
	Review JobTracker Nodes performance and perform appropriate OS and Application level tuning per CenturyLink Technology Solutions'	X	X	

Activity	Task	CenturyLink Basic Package	CenturyLink Basic + HBase Package	Customer
	and the Hadoop vendors standards, upon request of Customer			
	Create scheduled jobs and tasks with the Hadoop Cluster/Ecosystem			X
	Implement, Analyze, Modify or Remove Data Retention policies within the Hadoop Cluster/Ecosystem			X
	Manage the operational start and stopping of Hadoop services	X	X	
Monitoring	Monitor and Alert the customer on Raw and/or Usable disk space capacity being below a specified threshold	X	X	
	Monitor and Alert the customer and/or end users of Hadoop Cluster/Ecosystem, Long Running, Poorly Written or Performance Impacting Jobs and/or Tasks			X
	Monitor and Alert the customer to Hadoop Cluster/Ecosystem impacting events relating to Capacity, Availability or Performance	X	X	
	Provide Hadoop cluster name, server hostname and host component (NameNode, JobTracker, data node, etc) for manually created tickets as needed			X
Maintenance and Support	Data Gathering and Data Collection of Hadoop Cluster/Ecosystem Job and/or Task Activity	X	X	
	Data Gathering and Data Collection of Hadoop Cluster/Ecosystem related to Job and/or Task Failures	X	X	
	Data Gathering and Data Collection of Hadoop Cluster/Ecosystem related to Job and/or Task Resource Consumption/Time	X	X	
	Take applicable non-commercial action to restore availability of the Hadoop Cluster/Ecosystem	X	X	
	Data Gathering and Data Collection of Hadoop Cluster/Ecosystem general performance issues as reported by Customer or end users	X	X	
	Analysis of Hadoop Cluster/Ecosystem general performance issues as reported by the customer or end users specifically related to 3rd party integrations			X

Activity	Task	CenturyLink Basic Package	CenturyLink Basic + HBase Package	Customer
	Data Gathering, Data Collection, Analysis and Reporting of Hadoop Cluster/Ecosystem Hardware specific to physical hard disks, BIOS and Firmware settings/Versions and physical resource availability/consumption	X	X	
	Data Gathering, Data Collection, Analysis and Reporting of Hadoop Cluster/Ecosystem Hardware specific to Logical Resource Availability/Consumption, Driver and Kernel Settings/Versions and File System layout/Design	X	X	
	Modify and Manage the respective Customer User Accounts in the Hadoop Cluster/Ecosystem			X
	Terminate running Jobs or Tasks in the Hadoop Cluster/Ecosystem			X
	Add or Remove Data Nodes to the Hadoop Cluster/Ecosystem to add Capacity or Consolidate	X	X	
	Add or Remove Client Nodes to the Hadoop Cluster/Ecosystem to add Capacity or Consolidate	X	X	
	Add or Remove Control Nodes to the Hadoop Cluster/Ecosystem to add Capacity or Consolidate	X	X	
	Execute/Implement Minor Version or Security Updates/Patches to the Hadoop Cluster/Ecosystem upon customer request	X	X	
	Execute/Implement Major Version Updates/Patches to the Hadoop Cluster/Ecosystem			X
	Perform Tuning of the Hadoop Cluster/Ecosystem outside the scope of CenturyLink Technology Solutions' Standards or specific to a 3rd Party Integration			X
	Add or Remove Hadoop Cluster/Ecosystem Components post Installation/Burn-in, only as defined by this product.	X	X	
	Perform Hadoop Cluster/Ecosystem Component/Feature Migration from one Node to Another			X
	Execute/Implement Major, Minor or Security Updates/Patches to the Hadoop Cluster/Ecosystem specific to a 3rd Party Component or Software Integration (see Section			X

Activity	Task	CenturyLink Basic Package	CenturyLink Basic + HBase Package	Customer
	1.0 for patching for standard Service)			

Table 2.0 Monitors for the Service

Monitor Name	Hadoop Component	Description/Definition	Frequency
HADOOP_NativeHDFS	NameNode	Alarms when HDFS Filesystem threshold is exceeded, typical CENTURYLINK TECHNOLOGY SOLUTIONS traditional FS monitors are at 50%, 75% and 90%	60
HADOOP_LocalFS	NameNode	Alarms when Local Filesystem thresholds are exceeded	60
HADOOP_HDFS_NameNode_PROC	NameNode	Alarms HADOOP HDFS Server Process has died	60
HADOOP_HDFS_SecNameNode_PROC	Secondary NameNode	Alarms HADOOP HDFS Server Process has died	60
HADOOP_MAPRED_PROC	JobTracker	Alarms HADOOP MAPRED Server Process has died	60
HADOOP_HDFS_DataNode_PROC	DataNode	Alarms HADOOP HDFS DataNode Server Process has died	60
HADOOP_TASKTRACKER_DataNode_PROC	DataNode	Alarms HADOOP TaskTracker DataNode Server Process has died	60
HADOOP_NameNode_Port	NameNode	Checks that the Internal MetaData Transport is listening on port 8020	300
HADOOP_Backup_Port	Secondary NameNode	Checks that the Internal IPC Transport is listening on port 50100	300
HADOOP_DataNode_DFS_Port	DataNode	Checks that the Internal DFS Transport is listening on port 50010	300
HADOOP_DataNode_Recover_Port	DataNode	Checks that the Internal Recovery Transport is listening on port 50020	300

Monitor Name	Hadoop Component	Description/Definition	Frequency
HADOOP_JobTracker_Port	JobTracker	Checks that the Internal JobTracker Transport is listening on port 8021, 9001, or 8012 (JobTracker default ports are not well defined at this time)	300
HADOOP_TaskTracker_Port	DataNode	Checks that the Internal Transport is listening on IP Address 127.0.0.1:0***	300
HADOOP_NameNode_WUI_Port	Management Host	Checks that the Internal Web UI for HDFS NameNode Transport is listening on port 50070	300
HADOOP_DataNode_WUI_Port	Management Host	Checks that the Internal Web UI for HDFS NameNode Transport is listening on port 50075	300
HADOOP_SECOND_WUI_Port	Management Host	Checks that the Internal Web UI for HDFS NameNode Transport is listening on port 50090	300
HADOOP_BACKUP_WUI_Port	Management Host	Checks that the InternalWeb UI for HDFS Secondary NameNode Transport is listening on port 50105	300
HADOOP_JobTracker_WUI_Port	Management Host	Checks that the InternalWeb UI for JobTracker Transport is listening on port 50030	300
HADOOP_TaskTracker_WUI_Port	Management Host	Checks that the InternalWeb UI for JobTracker Transport is listening on port 50060	300
HADOOP_JVM_CPU	All	Alarms when CPU 80 % threshold is exceeded	60
HADOOP_JVM_CORE	All	Checks for Java Core or Heap Dumps	60
HADOOP_JDBC_Connections	Hbase	Alarms when DB Connection threshold is exceeded	60
HADOOP_JDBC_Connectivity	Hive and Pig	Alarms when the server cannot create an outbound connection.	60
HADOOP_JVM_CORE	All	Checks for Java Core or Heap Dumps	60
HADOOP_DataNode	All	Checks for dead DataNodes that have been	60

Monitor Name	Hadoop Component	Description/Definition	Frequency
_Remove		effectively removed from the Hadoop cluster	
HADOOP_BlackListed_Nodes	All	Checks for Hadoop 'Blacklisted' nodes present	60
HADOOP_CustomFailedDataFetcherJobs	All	Check for Failed Custom Data Fetcher Job Logs and Artifact Disk Usage Method : du -sh /mnt/hdgpfs/cluster/prod/permanentDirs/customDatafetcherFailedJobs	60

Definitions

Hadoop: Apache Hadoop is an open-source software framework for storage and large-scale processing of data sets on clusters of commodity hardware. Hadoop is an Apache top-level project being built and used by a global community of contributors and users.

HBase: Apache HBase, a key component of CDH, is a distributed, scalable data store that runs on top of HDFS. HBase is modeled after Google's Big Table and provides the ability to store data in massive tables (billions of rows / millions of columns) for fast, random access.

CenturyLink Service Center: The primary organization for resolving infrastructure issues that is staffed 24/7/365 to respond in a timely manner to incidents and requests pertaining to Customer IT infrastructure.

Enterprise Backup Data Recovery (BDR): Cloudera Enterprise BDR makes it easy to configure and manage backup disaster recovery policies for data stored in Cloudera Hadoop. With BDR you can centrally configure and manage disaster recovery workflows for files (HDFS) and metadata (Hive).

Impala: Cloudera Impala is a parallel processing SQL query engine that runs natively in Apache Hadoop. The Apache-licensed, open source Impala project combines modern, scalable parallel database technology with the power of Hadoop, enabling users to directly query data stored in HDFS and Apache HBase without requiring data movement or transformation. Impala is designed from the ground up as part of the Hadoop ecosystem and shares the same flexible file and data formats, metadata, security and resource management frameworks used by MapReduce, Apache Hive, Apache Pig and other components of the Hadoop stack.

Navigator: Cloudera Navigator is a fully integrated data management application for Apache Hadoop-based systems. It's designed to provide all of the capabilities required for administrators, data managers and analysts to secure, govern, and explore the large amounts of diverse data that land in Hadoop. As Hadoop deployments scale and data proliferates, the complexity around data management grows. More data from different sources gets loaded into the system. It has varying structures, schemas, access patterns and security restrictions. At the same time, there are different users that need to interact with the data in different ways. Administrators need to ensure that users have correct access and that security is properly enabled; data managers need to perform governance and audits of the

system for regulatory compliance; and analysts, data scientists, and other users need an easy way to find out what data is available to them, where it came from and what it looks like in order to use it effectively to gain insights that drive decisions.

Compiled: The compilation function turns source files into directly executable or intermediate objects. Not every project will require this function. While for simple programs the process consists of a single file being compiled, for complex software the source code may consist of many files and may be combined in different ways to produce many different versions.

Environment: The setting where software and other products are placed into operation for their intended uses by end users.

Flume: Apache Flume is a distributed, reliable, and available system for efficiently collecting; aggregating and moving large amounts of log data from many different sources to a centralized data store.

HDFS: Hadoop Distributed File System (HDFS) is a distributed file-system that stores data on commodity machines, providing very high aggregate bandwidth across the cluster.

Hive: The Apache Hive data warehouse software facilitates querying and managing large datasets residing in distributed storage. Hive provides a mechanism to project structure onto this data and query the data using a SQL-like language called HiveQL. At the same time this language also allows traditional map/reduce programmers to plug in their custom mappers and reducers when it is inconvenient or inefficient to express this logic in HiveQL.

Maintenance Windows: A period of time designated in advance by CenturyLink, during which preventive maintenance that could cause disruption of service may be performed. Current Scheduled Maintenance windows are:

- Americas: Saturday 00:00AM to 5:00AM; Sunday 00:00AM to 5:00AM
- EMEA: Saturday 02:00AM to 6:00AM
- APAC (Except Japan): Saturday 21:00 (GMT) AM to Sunday 01(GMT)
- Japan: Sunday 04:00 (JST) to 8:00 (JST)

Major Release: Major Releases (X.y.z) are vehicles for delivering major and minor feature development and enhancements to existing features. They incorporate all applicable error corrections made in prior Major Releases, Minor Releases, and Patch Releases. Software Provider typically has one Major Release per year.

Minor Release: Minor Releases (x.Y.z) are vehicles for delivering minor feature developments, enhancements to existing features, and defect corrections. They incorporate all applicable error corrections made in prior Minor Releases, and Patch Releases.

NameNode: The NameNode is the centerpiece of an HDFS file system. It keeps the directory tree of all files in the file system, and tracks where across the cluster the file data is kept. It does not store the data of these files itself.

Non-Production Environment: A product is still being used theoretically. Users, typically engineers, look for bugs or design flaws.

Oozie: Oozie is a workflow scheduler system to manage Hadoop jobs. It is a server-based Workflow Engine specialized in running workflow jobs with actions that run Hadoop MapReduce and Pig jobs. Oozie is implemented as a Java Web-Application that runs in a Java Servlet-Container.

Operating System: An operating system (OS) is software that manages computer hardware resources and provides common services for computer programs. The operating system is an essential component of the system software in a computer system. Application programs usually require an operating system to function.

Packaging: is the process of creating a meta-program that in turn automatically installs software across multiple computers. The meta-program typically includes a set of default properties for the applications it installs.

Patch Release: A patch is a small piece of software that is used to correct a problem with a software program or an operating system. Patches are often called "Patch Updates" and are Critical or Security related.

Pig: Apache Pig is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs. The salient property of Pig programs is that their structure is amenable to substantial parallelization, which in turns enables them to handle very large data sets. At the present time, Pig's infrastructure layer consists of a compiler that produces sequences of Map-Reduce programs, for which large-scale parallel implementations already exist. Pig's language layer currently consists of a textual language called Pig Latin, which allows easy programming of MapReduce jobs.

Production Environment: A production environment can be thought of as a real-time setting where programs are run and hardware setups are installed and relied on for organization or commercial daily operations.

Secondary NameNode: The Secondary NameNode (SNN) is an assistant daemon for monitoring the state of the cluster, Like the NameNode, Each cluster has one SNN, and it typically resides on its own machine.

Service Level Agreement: A service-level agreement (SLA) is a document describing the level of service expected by a customer from CenturyLink, laying out the metrics by which that service is measured, and the remedies or penalties, if any, should the agreed-upon levels not be achieved.

SLA Credit: Service Level Agreement Credits are refunds given by CenturyLink to a Customer if the service falls below a contractually agreed service levels. See Section 1.0 for the SLA agreement associated with this service

Sqoop: Apache Sqoop is a tool designed for efficiently transferring bulk data between Apache Hadoop and structured data stores such as relational databases. You can use Sqoop to import data from external structured data stores into the Hadoop Distributed File System (HDFS) or related systems such as Hive and HBase. Conversely, you can use Sqoop to extract data from Hadoop and export it to external structured data stores such as relational databases and enterprise data warehouses.

Software Binary: A binary file is a file whose content must be interpreted by a program or a hardware processor that understands in advance exactly how it is formatted.

Software Provider: The third party that makes and sells the Software products described in section 1.0.

Statement of Work: A statement of work (SOW) is a formal document that captures and defines the work activities, deliverables, and timeline a vendor must execute in performance of specified work for a client. The SOW usually includes detailed requirements and pricing, with standard regulatory and governance terms and conditions.

Third Party Software: A reusable software component developed to be either freely distributed or sold by an entity other than the original vendor of the development platform.

ZooKeeper: ZooKeeper is a centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services. All of these kinds of services are used in some form or

another by distributed applications. Each time they are implemented there is a lot of work that goes into fixing the bugs and race conditions that are inevitable.

Upgrades: Upgrades mean a Major Release or Minor Release of the software.