**Oozie Questions & Answers**

**Important Points:**  
Apache Oozie is a server-based workflow scheduling system to manage Hadoop jobs.Workflows in Oozie are defined as a collection of control flow and action nodes in a directed acyclic graph.

Control flow nodes define the beginning and the end of a workflow (start, end, and failure nodes) as well as a mechanism to control the workflow execution path (decision, fork, and join nodes).   
  
**1)What is Oozie?**   
**Answer:** Oozie is a workflow scheduler for Hadoop Oozie allows a user to create Directed Acyclic Graphs of workflows and these can be ran in parallel and sequential in Hadoop.It can also run plain java classes, Pig workflows and interact with the HDFS .It can run jobs sequentially and in parallel.   
  
**2)Why use oozie instead of just cascading a jobs one after another?**   
**Answer:** Major Flexibility :Start ,stop ,re-run and suspend  
Oozie allows us to restart from failure  
  
**3)How to make a workflow?**   
**Answer:** First make a Hadoop job and make sure that it works Make a jar out of classes and then make a workflow.xml file and copy all of the job configuration properties in to the xml file.  
Input files  
Output files  
Input readers and writers  
mappers and reducers  
job specific arguments  
job.properties  
  
**4)What are the properties that we have to mention in .Properties?**   
**Answer:** Name Node  
Job Tracker  
Oozie.wf.application.path  
Lib Path & Jar Path  
  
**5)What is application pipeline in Oozie?**   
**Answer:** It is necessary to connect workflow jobs that run regularly, but at different time intervals. The outputs of multiple subsequent runs of a workflow become the input to the next workflow. Chaining together these workflows result it is referred as a data application pipeline.   
**6)How to run Oozie?**   
**Answer:** $ oozie job -oozie http://172.20.95.107:11000(oozie server node)/oozie -configjob.properties -run  
This will give the job id.  
To know the status: $ oozie job -oozie http://172.20.95.107:11000(oozie server node)/oozie -info job id  
  
**7)What are all the actions can be performed in Oozie?**   
**Answer:** Email Action  
Hive Action  
Shell Action  
Ssh Action  
Sqoop Action  
Writing a custom Action Executor  
  
**8)Why we use Fork and Join nodes of oozie?**   
**Answer:** A fork node splits one path of execution into multiple concurrent paths of execution.  
A join node waits until every concurrent execution path of a previous fork node arrives to it.  
The fork and join nodes must be used in pairs. The join node assumes concurrent execution paths are children of the same fork node.  
  
**9)What is Decision Node in Oozie?**   
**Answer:** Decision Nodes are switch statements that will run different jobs based on the outcomes of an expression.   
  
**10)Mention Some Features Of Oozie?**   
**Answer:** Oozie has client API and command line interface which can be used to launch, control and monitor job from Java application.  
Using its Web Service APIs one can control jobs from anywhere.  
Oozie has provision to execute jobs which are scheduled to run periodically.  
Oozie has provision to send email notifications upon completion of jobs.  
  
**11)Explain Need For Oozie?**   
**Answer:** With Apache Hadoop becoming the open source de-facto standard for processing and storing Big Data, many other languages like Pig and Hive have followed - simplifying the process of writing big data applications based on Hadoop.  
Although Pig, Hive and many others have simplified the process of writing Hadoop jobs, many times a single Hadoop Job is not sufficient to get the desired output. Many Hadoop Jobs have to be chained, data has to be shared in between the jobs, which makes the whole process very complicated.  
**12)What Are The Alternatives To Oozie Workflow Scheduler?**   
**Answer:** Azkaban is a batch workflow job scheduler  
Apache NiFi is an easy to use, powerful, and reliable system to process and distribute data.  
Apache Falcon - Feed management and data processing platform  
  
**13)Explain Briefly About Oozie Bundle?**   
**Answer:** Oozie Bundle is a higher-level oozie abstraction that will batch a set of coordinator applications. The user will be able to start/stop/suspend/resume/rerun in the bundle level resulting a better and easy operational control.  
More specififcally, the oozie Bundle system allows the user to define and execute a bunch of coordinator applications often called a data pipeline. There is no explicit dependency among the coordinator applications in a bundle. However, a user could use the data dependency of coordinator applications to create an implicit data application pipeline.  
Oozie executes workflow based on:  
Time Dependency(Frequency)  
Data Dependency  
  
**14)What Is Application Pipeline In Oozie?**   
**Answer:** It is necessary to connect workflow jobs that run regularly, but at different time intervals. The outputs of multiple subsequent runs of a workflow become the input to the next workflow. Chaining together these workflows result it is referred as a data application pipeline.   
  
**15)Why Oozie Security?**   
**Answer:** User are not allowed to alter job of another user  
Hadoop does not support the authentication of end user  
Oozie has to verify and confirms its user before transferring the job to Hadoop