

Name			Self rating score	0
Total years of experience			Assesment score	
Data engineering experience				
Prepared by		Prithiviraj Damodaran		
		<i>Items highlighted in grey are "design thinking" scenarios that candidates will be given to work on.</i>		
Environments		Competencies	Response / Self-Rating	Response type
		Hadoop distributions worked on		HDP, CDH, MapR, AWS, MS Azure, IBM BigInsights
		Cloud Service providers exposed to		AWS , GCE
Programming languages		<i>Java</i>		0=Beginner, 1=Proficient, 2=Expert
		<i>Scala</i>		0=Beginner, 1=Proficient, 2=Expert
		<i>Python</i>		0=Beginner, 1=Proficient, 2=Expert
		<i>R</i>		0=Beginner, 1=Proficient, 2=Expert
		<i><Add></i>		
		<i><Add></i>		
Data Capture /Collection		<i>Event Ingestion</i>		
		<i>Data Integration from multiple legacy data sources</i>		
		<i>Event sourcing from multiple data sources</i>		
		<i>Expsoure to libraries, tools and frameworks, if any</i>		Kafka, AWS Kinesis, Flume, Flafka, Kafka connect, Gobblin, SQOOP
		<i><Add></i>		
		<i><Add></i>		
Data processing	Batch Processing	<i>MapReduce (MR)</i>		0=Not worked, 1=Built-prototype, 2=Built-production solutions
		<i>MR Abstractions - PIG, Cascading/JCascalog/Scalding</i>		0=Not worked, 1=Built-prototype, 2=Built-production solutions
		<i>Spark</i>		0=Not worked, 1=Built-prototype, 2=Built-production solutions
		<i><Add></i>		0=Not worked, 1=Built-prototype, 2=Built-production solutions
		<i>How to do you choose a framework for problem</i>		
	Micro-Batch / Stream / Hybrid processing	<i>Storm, Trident, DRPC</i>		0=Not worked, 1=Built-prototype, 2=Built-production solutions
		<i>Spark Streaming</i>		
		<i>Flink</i>		0=Not worked, 1=Built-prototype, 2=Built-production solutions
		<i>Google cloud dataflow a.k.a Apache Beam</i>		0=Not worked, 1=Built-prototype, 2=Built-production solutions
		<i>Kafka Streams</i>		0=Not worked, 1=Built-prototype, 2=Built-production solutions
		<i>S4, Samza</i>		0=Not worked, 1=Built-prototype, 2=Built-production solutions
		<i><Add></i>		0=Not worked, 1=Built-prototype, 2=Built-production solutions
		<i>How do you choose a framework for problem ? what are the factors to be considered ?</i>		
		<i><Add></i>		
<i><Add></i>				

		<Add>		
Data Storage		NoSQL		
		<Add>		
		How do you choose a NoSQL data store for a given usecase ?		Cassandra, HBASE, MongoDB, Neo4J
		Writing data to HDFS EDL implementation experiences Using HDFS-as-a-sink in frameworks HDFS file formats Compression "Updating" HDFS files with periodic changes		
		NewSQL		
		RDBMS principles, modelling		
		SQL-On-Hadoop variants		
		MR based		
		<i>Hive on MR</i>		0=Not worked, 1=Built-prototype, 2=Built-production solutions
		Special purpose DAG engines based		
		<i>Hive on Tez</i>		0=Not worked, 1=Built-prototype, 2=Built-production solutions
		<i>Hive on Spark</i>		0=Not worked, 1=Built-prototype, 2=Built-production solutions
		MPP-on-hadoop / distributed query engines		
		<i>Impala</i>		0=Not worked, 1=Built-prototype, 2=Built-production solutions
		<i>Presto</i>		0=Not worked, 1=Built-prototype, 2=Built-production solutions
		<i>Drill</i>		0=Not worked, 1=Built-prototype, 2=Built-production solutions
		BigQuery		0=Not worked, 1=Built-prototype, 2=Built-production solutions
		In-memory DAG		0=Not worked, 1=Built-prototype, 2=Built-production solutions
		<i>Sparksql</i>		0=Not worked, 1=Built-prototype, 2=Built-production solutions
		Connecting BI/Viz tools to Bigdata		Tableau, Qlikview, Microstrategy, BO, Cognos
		HDFS, Hive, HBASE direct		
		Via Spark		
		<i>Expsoure tools and frameworks, if any</i>		
		<Add>		
		How to do you choose a Sql-on-hadoop option for a problem ? what are the factors to be considered ?		
		<Add>		
		<Add>		
Data Viz		Experiences in data visualisation using BI tools or custom		
		<Add>		
		<Add>		
Misc		Any other relevant tools, frameworks, managed services and libraries		
		Any non-apache/commercial tools experiences in any of the above spaces ?		
		Experiences in AWS and GCE managed services.		
		Good to have		
		Unix commandline		
		<Add>		