05: Q37-Q41 – Data lake & metadata interview questions & answers

Posted on August 6, 2016 by Arulkumaran Kumaraswamipillai

Q37. What is a Data Lake?

A37. A **data lake** is a storage repository that holds a vast amount of structured, semi-structured, and unstructured **raw data** in its native format (aka pristine condition). The data structure and requirements are not defined until the data is needed. You can also call it a “raw data zone”.

In the Hadoop world, this raw data can be can ingested from the various sources via Apache Flume agents, Kafka, Apache NiFi, etc. The sources can be files, messages from a MOM (Message Oriented Middle-ware) like Webspehere MQ, etc.

600+ Full Stack Java/JEE Interview Q&As ♥Free ♦FAQs

open all | close all

- Ice Breaker Interview
- Core Java Interview C
- JEE Interview Q&A (3
- Pressed for time? Jav
- SQL, XML, UML, JSC
- Hadoop & BigData Int
  - ♥ 01: Q1 – Q6 Had
  - 02: Q7 – Q15 Hadc
  - 03: Q16 – Q25 Hac
  - 04: Q27 – Q36 Apa
  - 05: Q37 – Q50 Apa
  - 05: Q37-Q41 – Dat
  - 06: Q51 – Q61 HBa
  - 07: Q62 – Q70 HDI
- Java Architecture Inte
- Scala Interview Q&As
- Spring, Hibernate, & M
- Testing & Profiling/Sa
- Other Interview Q&A f
- ▶Free Java Interviev

Q38. How does it differ from the data warehouse?

A38.

Data Warehouse	Data LAKE
Used by business users	Used by data scientists & analysts
Structured & processed data.	Structured, Semi-structured (e.g. XML), and unstructured (e.g. PDFs, word, images, etc) in raw. This unstructured nature of raw data gives better agility for data scientists & developers to re-configure their models & query on the fly. The raw data can be manipulated in a variety of ways.
schema-on-write. Before we can load data into a data warehouse, we first need to give it some shape and structure.	**schema-on-read (aka late binding)**. In a data lake, you just load in the raw data, as-is, and when you're ready to use the data with a certain access pattern, you give it a shape and structure.
Expensive for large data volumes as run on specialized hardware.	Low cost storage as Hadoop runs on commodity hardware & Hadoop is open-source without any licensing fees. You also have community support.

Q39. How would you prevent a “data lake” quickly becoming a “data swamp”?
A39. You need to have an effective “data management” layer to prevent a “data lake” becoming a “data swamp”. The data management layer involves:

1) Extensible metadata registry: that provides data discovery & data lineage management functions. For example, “Cloudera Navigator metadata component”, “Loom” from Revelytix, etc. The metadata needs to capture security classification, data source info, created timestamp info, time to live info, source system, etc.

2) Tracking data transformations and lineage. Recording all the input and output transformation processes via MapReduce or Spark jobs.

3) RESTful APIs to discover metadata & data from other enterprise platforms & tools.

4) When you move data from the raw zone to “redefined/user” zones where a structure is applied on read to be used by data scientists, it is imperative to apply **data quality management**. Duplicate & bad data needs to be re-mediated & cleansed.

5) Develop repeatable and controlled processes to ingest and transform data into Hadoop.

6) Implement proper security, auditing & monitoring. E,g. Cloudera Navigator, Knox (Central authentication mechanism for Hadoop), Access control with LDAP, encryption of sensitive data, proper logging, etc.

Metadata, transformation lineage & quality of data are imperative to ensure that “data lake” does not become a “data swamp”.

Q40. How does the Cloudera Navigator Metadata Server extract metadata from the entities managed by Cloudera

Manager?

A40.

1) HDFS at the next scheduled run after an HDFS checkpoint. **Checkpointing** is a process of maintaining & persisting filesystem metadata in HDFS. It is crucial for NameNode recovery & restart. The filesystem metadata is stored in “fsimage” and the “edit log”. The “fsimage” is a file that represents a point-in-time snapshot of the filesystem’s metadata. The fsimage file format is very efficient to read, but not for making small incremental updates like renaming a single file. Hence, the “edit log” is used for recording the updates. This way, if the NameNode crashes, it can restore its state by first loading the “fsimage”, and then replaying all the operations from the “edit log”.

2) Hive/Sqoop 1 extracts table metadata from the **Hive Metastore server**.

3) MapReduce/Pig extracts the job metadata from the JobTracker.

4) YARN/Pig extracts from the Job History Server.

Q41. How would you go about “organizing your data lake” in HDFS?

A41. Firstly you can have 3 main directories.

1. Raw or Staging: to host all the original source files as they get ingested into the data lake. Each source system should have its own directory. For example:

```
1
2 /data/raw/client1/day1
3 /data/raw/client2/day2
4 /data/stage/client1/20150523
5
```

2. Cleansed or redefined: The raw or staged data needs to go through basic quality check. For example, trade files need to have valid symbol, price, and volume info. The cleansing

also include de-duplication of data. The cleansed data can be grouped by subject domains like finance, logs, sales, etc.

```
1
2 /data/cleansed/client1/finance/day1
3 /data/cleansed/client2/sales/day2
4 /data/redefined/client1/logs/20150523
5
```

3. Summarised or user defined: is the precomputed & optimized data that is used by the data scientists or analysts to be used for querying & reporting. For example, quarterly sales report, etc.

```
1
2 /data/summarised/client1/sales/quarterly
3 /data/summarised/client2/sales/monthly
4 /data/user_defined/client1/bugs_report/2016_June
5
```

In order to keep track of all the files and their corresponding directories, you need to maintain a repository of meta data that can be indexed (E.g. Apache Solr). The metadata info like security classification, data source info, ingestion timestamp , time to live, source system, file size, structure of the file, key columns, etc can be stored & indexed. An Oozie workflow can be created to scan the data files in HDFS, and updates or builds the SOLR index. The file meta data can also be maintained in an HBase/Hive database. We can query the metadata via Hive/Pig.

Popular Posts

♦ 11 Spring boot interview questions & answers

828 views

♦ Q11-Q23: Top 50+ Core on Java OOP Interview Questions & Answers

768 views

18 Java scenarios based interview Questions and Answers

401 views

001A: ♦ 7+ Java integration styles & patterns interview questions & answers

389 views

01b: ♦ 13 Spring basics Q8 – Q13 interview questions & answers

296 views

♦ 7 Java debugging interview questions & answers

293 views

01: ♦ 15 Ice breaker questions asked 90% of the time in Java job interviews with hints

286 views

♦ 10 ERD (Entity-Relationship Diagrams) Interview Questions and Answers

280 views

♦ Q24-Q36: Top 50+ Core on Java classes, interfaces and generics interview questions & answers

240 views

001B: ♦ Java architecture & design concepts interview questions & answers

202 views

Bio | Latest Posts

Arulkumaran Kumaraswamipillai

Mechanical Eng to freelance Java developer in 3 yrs. Contracting since 2003, and attended 150+ Java job interviews, and often got 4 - 7 job offers to choose from. It pays to prepare. So, published Java interview Q&A books via Amazon.com in 2005, and sold 35,000+ copies. Books are outdated and replaced with this subscription based site.**945+** paid members. join my LinkedIn Group. **Reviews**

About Arulkumaran Kumaraswamipillai

Mechanical Eng to freelance Java developer in 3 yrs. Contracting since 2003, and attended 150+ Java job interviews, and often got 4 - 7 job offers

to choose from. It pays to prepare. So, published Java interview Q&A books via Amazon.com in 2005, and sold 35,000+ copies. Books are outdated and replaced with this subscription based site.**945+** paid members. join my LinkedIn Group. **Reviews**

‹ YAML with Java using the SnakeYaml library tutorial

03: Create or append a file to HDFS – Hadoop API tutorial ›

Posted in Hadoop & BigData Interview Q&A, member-paid

Empowers you to open more doors, and fast-track

Technical Know Hows

☀ Java generics in no time ☀ Top 6 tips to transforming your thinking from OOP to FP ☀ How does a HashMap internally work? What is a hashing function?
☀ 10+ Java String class interview Q&As ☀ Java auto un/boxing benefits & caveats ☀ Top 11 slacknesses that can come back and bite you as an experienced Java developer or architect

Non-Technical Know Hows

☀ 6 Aspects that can motivate you to fast-track your career & go places ☀ Are you reinventing yourself as a Java developer? ☀ 8 tips to safeguard your Java career against offshoring ☀ My top 5 career mistakes

Prepare to succeed

☀ Turn readers of your Java CV go from “Blah blah” to “Wow”? ☀ How to prepare for Java job interviews? ☀ 16 Technical Key Areas ☀ How to choose from multiple Java job offers?

Select Category

© Disclaimer

↑