Designing Data Processing Systems: Exam Guide Review

Storage

Selecting the appropriate storage technologies.	Mapping storage systems to business requirements. Data modeling **Tradeoffs involving latency, throughput, and transactions** Distributed systems Schema design

Tip: Be familiar with the common use cases and qualities of the different storage options. Each storage system or database is optimized for different things -- some are best at atomically updating the data for transactions. Some are optimized for speed of data retrieval but not for updates or changes. Some are very fast and inexpensive for simple retrieval but slow for complex queries.

Pipelines

Designing data pipelines.	**Data publishing and visualization** Batch and streaming Online (interactive) vs. batch predictions Job automation and orchestration

Tip: An important element in designing the data processing pipeline starts with selecting the appropriate service or collection of services.

Tip: AI Platform Notebooks, Google Data Studio, BigQuery all have interactive interfaces. Do you know when to use each?

Processing Infrastructure

Designing a data processing solution.	Choice of infrastructure System availability and fault tolerance Use of distributed systems Capacity planning Hybrid cloud and edge computing Architecture options **At least once, in-order, and exactly once event planning**

Tip: Pub/Sub and Dataflow together provide once, in-order, processing of possibly delayed or repeated streaming data.

Tip: Be familiar with the common assemblies of services and how they are often used together: Dataflow, Dataproc, BigQuery, Cloud Storage, and Pub/Sub.

Migration

Migrating data warehousing and data processing.	Awareness of current state and how to migrate design to a future state **Migrating from on-premise to cloud** Validating a migration

Tip: Technologically, Dataproc is superior to Open Source Hadoop, and Dataflow is superior to Dataproc. However, this does not mean that the most advanced technology is always the best solution. You need to consider the business requirements. The client might want to first migrate from the data center to the cloud. Make sure everything is working (validate it). And only after they are confident with that solution, to consider improving or modernizing.