Simple Application

Working Backwards

* Start with the end user’s needs, not from where your data is coming from
* What sort of access patterns do you anticipate from your end users?
  + Analytical queries that span large date ranges”
  + Huge amounts of small transactions for very specific rows of data?
  + Both?
* What Availability do these users demand?
* What consistency do these users demand?

Requirments

* How big is your big data?
* How much internal infrastructure and expertise is available?
  + AWS?
  + System you already know fit the bill?
* What about data retention?
  + Do you need to keep data around forever, for auditing?
  + Or do you need to purge it often, for privacy?
* What about security?
  + check with legal

Latency

* How quickly do end users need to get a response?
  + milliseconds? then something like Hbase or cassandra will be needed?

Timeliness

* Can queries be based on day-old data? minute-old?
  + Oozie-scheduled jobs in Hive/Pig/Spark etc may cut it
* Or must it be near-real-time?
  + Use spark streaming/storm/flink with Kafka or Flume

Future proofing.

Cheat to win - Utilize the tools that you can leverage on.