Quiz 1 - Chapter 1

1. Name and describe one of the desired properties of a big data system. (1 sentence)

    a. Robustness and fault tolerance - Systems need to behave correctly despite machine failure and errors.
    b. Low latency reads and updates - A big data system should serve serve data and answer queries within some acceptable threshold of time.
    c. Scalability - A big data system should be scalable; that is, able to maintain performance in the face of increased data or load by adding resources to the system.
    d. Generalization - A general big data system can support a wide range of applications.
    e. Extensibility - Extensible systems allow functionality to be added with a minimal development cost.
    f. Ad hoc queries - A big data system should support the ability to perform new queries with minimal work as this form of analysis gives opportunities for business optimization and new applications.
    g. Minimal Maintenance - A big data system should be able to stay up-and-running with as little maintenance as possible.
    h. Debuggability - A big data system must provide the information necessary to debug errors as they occur.

2. Name and describe each layer of the Lambda Architecture. (3 sentences)
    a. The Batch Layer stores an immutable, constantly growing master dataset and must be able to compute arbitrary functions on that dataset. It precomputes batch views to help support low latency queries.
    b. The Serving Layer is a specialized distributed database that serves that batch views produced by the Batch Layer.
    c. The Speed Layer ensures that new data is represented in query functions by updating realtime views incrementally. It resolves the lag produced by the Batch Layer.

3. Describe what you learned in the Spotify blog post. (3 sentences)

There are many acceptable answers to this question. I was just looking for something that would demonstrate a student read the article. One such possible answer is:

The Spotify blog post discussed Spotify's event delivery system, which is responsible for transporting each user interaction from within the Spotify application to Spotify's data centers and storing these interactions. All of these interactions are delivered to a Hadoop cluster that Spotify manages. Although Spotify had to rebuild their event delivery system because their first version was very complex and had several issues, including sending unstructured data over the network, it was still able to handle about 700,000 events per second.