Questions for the written part of the literature seminar

These questions should be answered as a team. The articles in Studium serve as a starting point, you are expected to follow references in those articles and additionally look for relevant literature on your own. At least two references in your essays should be to a scientific paper not provided in the student portal. Your essays should contain references to scientific literature, in a consistent referencing format.

Q1: Discuss and contrast the characteristics of batch vs. streaming applications, both from the problem point of view and from a technology point of view.

Q2: Discuss in what way the role of data types and formats place on the technological solutions. Are there specialized tools that achieve high-performance and high usability for specific formats?

Q3: Try to, from a technological point of view, relate the following tools to each other, both historically and in the technological problems they have addressed. Is there some logical progression in a seemingly fragmented ecosystem?

Apache Hadoop, HDFS, Hbase, BigTable, Cassandra, Twitter Storm, Twitter Heron, Apache Spark, Apache Kafka, Apache Flink (of course, include more tools if you find things that interests you)

Q4: Why do you think NoSQL alternative such as MongoDB or Apache Cassandra has gained so much in popularity?

Q5: Scalability is an important concept in distributed systems. Discuss this both from a theoretical basis and from a technological (i.e. what are current ways to achieve this). What is the difference between reactive and proactive autoscaling, and what are some of the current approaches/proposals to achieve it?