*Batch data* are often processed in intervals and involve a large grouping of records, and data warehouses are often refreshed on a set interval, using batch processes. An example would be a financial warehouse refreshed nightly and that is when any new records would be added. Streaming data are quite different. They are constantly being generated and need to be processed as soon as possible. An example of streaming data would be a social media feed for a retail company or Internet of Things (IoT) for robotic assembly lines.

For this assignment, you will evaluate the characteristics of streaming data versus batch data pipelines.

* What are the major differences between streaming data and batch data pipelines that you see?
* What kind of situation would a company be inclined to use streaming data over batch data?
  + Provide a real-world example of streaming data and an example of batch data.

**Differences between streaming data and batch data pipelines**

|  |  |
| --- | --- |
| **Streaming Data Pipelines** | **Batch Data Pipelines** |
| Constant arrival of small amounts of data | Data arrives in larger chunks |
| Analysis happens as the data is generated | Analysis happens after the data is generated |
| Data is stored in the data warehouse after it has been analyzed | Data is usually stored in a data warehouse before analysis occurs, ETL process to ready data for the warehouse |
| Algorithms that compute on 1 element or a small window of elements are necessary since data is constantly arriving (size unbounded) | Algorithms can compute over the entire batch of data |
| Fast and simple computations to keep up with the stream of data | Computations can be more complex and take longer since there is no time constraint |
| Producers of data and consumers of data are decoupled. An intermediary broker or buffer system holds data that consumers can then access at their own pace. | No need for a buffer to hold data because it is static. Once it is created, it is stored at rest and accessed to perform analytics. |
| The size and frequency of streams of data is dynamic and hard or impossible to predict. | The size of the data is predetermined and constant. |

(Coursera, n.d.; Simplilearn, 2017; *What Is Streaming Data?*, n.d.)

**Use Case for Streaming Data:**

Companies may be inclined to use streaming data when the data is **time sensitive** and **continuously generated**. For example, the **Netflix movie/recommendation engine** relies on streaming data. Data is constantly coming into the system: what movies I have watched, if I clicked the “like” or “not for me” options, what other people who have watched those same movies have watched/liked, etc. That data is useful if it can be instantaneously analyzed to provide me with recommendations of what to watch next so that I stay on their platform. This is an example of time sensitive data where the influx of data has no set start or end point.

**Use Case for Batch Data:**

Batch data is more useful for companies when data only needs to be managed periodically. **Storing data in a data** warehouse can happen at the end of each day, rather than for every event. HR systems use batch data to **calculate payroll** periodically (biweekly, monthly, etc.) (*What Is Batch Processing?*, n.d.).

**References**

Coursera (Director). (n.d.). *Why is Streaming Data different? - Working With Data Models* [Video recording]. Retrieved October 15, 2024, from https://www.coursera.org/lecture/big-data-management/why-is-streaming-data-different-Wfb3n

Simplilearn (Director). (2017, July 25). *Kafka Tutorial | Apache Kafka Tutorial For Beginners | Kafka Architecture |What Is Kafka|Simplilearn* [Video recording]. https://www.youtube.com/watch?v=U4y2R3v9tlY

*What is Batch Processing? - Batch Processing Systems Explained - AWS*. (n.d.). Amazon Web Services, Inc. Retrieved October 15, 2024, from https://aws.amazon.com/what-is/batch-processing/

*What is Streaming Data? - Streaming Data Explained - AWS*. (n.d.). Amazon Web Services, Inc. Retrieved October 15, 2024, from https://aws.amazon.com/what-is/streaming-data/