Санкт-Петербургский государственный университет

Работа допущена к защите

зав. кафедрой

_____

«____» _____ 2021 г.

Курсовая работа

Тема: **Сравнительный анализ систем загрузки больших данных**

Направление: 010400 (01.03.02) – Прикладная математика и информатика

Выполнил студент гр. 323 _____ Мамаев Владислав Викторович

Научный руководитель,

к. ф.-м. н., _____ Благов Михаил Валерьевич

Санкт-Петербург – 2021

Оглавление

1. Введение

Базы данных используются повсеместно, при этом системы, для которых они предназначены, можно разделить на 2 класса:

- Online Transaction Processing (OLTP) – системы обработки транзакций в реальном времени. Такие системы используются для операционной деятельнос предприятий.
- Online Analytical Processing (OLAP) – системы аналитической обработки в реальном времени. К такими системам относятся системы поддержки принятия решений, инструменты business intelligence, системы анализа данных.

OLAP системы содержат информацию из OTLP систем, при обычно OLAP системы обычно функционирует независимо от OLTP систем по следующим причинам: [1]

- Реляционная схема данных, обычно используемая в OLTP системах, не эффективна для OLAP нагрузки.
- OLAP нагрузка на OLTP систему может вызвать проблемы с производительнос транзакций.
- С увеличением размера хранимых данных, возникают ограничения на используемые технологии и существенно увеличивается стоимость хранения данных в системах не предназначенных только для OLAP.
- Необходимость доступа к данным из разных OLAP систем и возможности ограничения доступа к данным

Традиционно, данные попадают в OLAP систему из OLTP системы при помощи Extraction-Transformation-Loading (ETL) программ, запускаемых с некоторо

периодичностью. ETL процессы могут занимать достаточно много времени, и это создает задержку появления данных в OLAP системе.

Для того чтобы уменьшить задержку можно вместо переодичных ETL процессов использовать потоковую обработку. Подобная архитектура интеграции данных описана в [2] и представленна на рисунке TODO. Каждое изменение данных в OLTP системе записывается в очередь сообщений, а некоторая программа непрерывно читает сообщения и обновляет данные в аналитическом хранилище. Использование непрерывных обновлений уменьшает задержку, но требует специальн инструментов для обеспечения целостности данных — традицонно используемые хранилища для большого объема данных, такие как Hadoop, работают по принципу write-once и не поддерживают обновлений.

В данной работе будет проведено сравнение таких инструментов: Apache Hudi, Delta Lake, Apache Iceberg. Сравнение будет произведино по следующим пунктам:

- Время задержки доставки данных.
- Пропускная способность.
- Удобство доступа к данным — наличие интеграций с другими инструментами для работы с большими данными и совместимость.
- Безопасность.

2. Основная часть

[3]

2.1. Основная часть

Список литературы

1. *Conn S. S.* OLTP and OLAP data integration: a review of feasible implementation methods and architectures for real time data analysis // Proceedings. IEEE SoutheastCon, 2005. — 2005. — С. 515—520. — DOI: 10.1109/SECON.2005.1423297.

2. *Tho M. N.*, *Tjoa A. M.* Zero-latency data warehousing for heterogeneous data sources and continuous data streams // 5th International Conference on Information Integrationand Web-based Applications Services. — 2003. — С. 55—64.

3. Data Ingestion for the Connected World. / J. Meehan [и др.] // CIDR. — 2017.