1. Спецкурс программы аспирантуры, полугодовой: **Базы данных: дополнительные главы.**

2. Преподаватели: доц. С.Т.Главацкий, н.с. И.Г.Бурыкин.

3. Аннотация курса: базы данных в памяти: методы кодирования (словарное, префиксное, кластерное, разреженное, непрямое, дельта, длин серий), размещение данных в оперативной памяти (построчное, по столбцам, гибридное), секционирование (по группам атрибутов, по кортежам, по диапазону значений, циклическое, с хешированием, семантическое), архивирование и старение данных, операторы манипулирования данными, стратегии материализации, индексы, горячее резервирование и k-безопасность, бизнес-логика в базе данных;

NoSQL базы данных: теорема CAP, размещение данных (документно-ориентированное, на основе графов, "ключ-значение", семейства колонок);

распределенный реестр: консенсус модели, блокчейн (криптографические хэш-функции, хэш-указатели, цепочки блоков и деревья Меркля).

4. Тематическое содержание курса:

|  |  |
| --- | --- |
| **Тема 1.** | Прошлое, настоящее и будущее корпоративных приложений. |
| **Тема 2.** | Методы хранения баз данных, словарное кодирование. |
| **Тема 3.** | Сжатие данных: словарное, префиксное, кластерное, разреженное, непрямое, дельта кодирование и кодирование длин серий. |
| **Тема 4.** | Размещение данных в оперативной памяти и секционирование. |
| **Тема 5.** | Архивирование и старение данных, мульти-температурное управление данными, фактические и исторические данные. |
| **Тема 6.** | Операторы манипулирования данными (DML): изменение содержимого базы данных. |
| **Тема 7.** | Операторы манипулирования данными (DML): выборка и агрегатные функции. |
| **Тема 8.** | Стратегии материализации и алгоритмы соединения таблиц. |
| **Тема 9.** | Архитектура базы данных в памяти. |
| **Тема 10.** | Индексы, кеширование для агрегатов, резервные копии и восстановление. |
| **Тема 11.** | Горячее резервирование и k-безопасность, реплики, бизнес-логика в базе данных. |
| **Тема 12.** | Параллельная обработка данных, закон Амдала. |
| **Тема 13.** | Конфликты и изолированность транзакций. |
| **Тема 14.** | Прогнозирование и интеллектуальный анализ текстов в базах данных в памяти. |
| **Тема 15.** | Базы данных NoSQL, теорема CAP, размещение данных. |
| **Тема 16.** | Базы данных и технология распределенного реестра, консенсус модели. |
| **Тема 17.** | Блокчейн и криптовалютные технологии. |

5. Типовые контрольные задания или иные материалы, необходимые для оценки результатов обучения, характеризующих этапы формирования компетенций.

1. Корпоративные приложения: современные требования, проблемы в аппаратном и программном обеспечении и способы их решения, характеристики современных корпоративных приложений, OLTP и OLAP системы.

2. Индексы в In-Memory базе данных, инвертированный индекс, эффективность инвертированного индекса при словарном кодировании.

3. Методы хранения баз данных: секционирование, мульти-температурное управление данными, фактические и исторические данные.

4. Сжатие данных в In-Memory базах данных: префиксное кодирование (Prefix Encoding), разреженное кодирование (Sparse encoding), непрямое кодирование (Indirect encoding).

5. Операторы манипулирования данными в In-Memory базе данных: вставка, обновление, удаление, "insert only"-подход.

6. Построчное и поколоночное размещение данных: сравнение подходов при реконструкции кортежей.

7. Оператор выборки в In-Memory базе данных: стратегии материализации, агрегатные функции.

8. Словарное кодирование в In-Memory базе данных, понятие энтропии.

9. In-Memory СУБД: архитектура, понятие дифференциального буфера и процесс слияния, кеш для агрегатов.

10. Построчное и поколоночное размещение данных: сравнение подходов при полном сканировании таблицы, шаговом доступе и полном сканировании столбца.

11. In-Memory СУБД: ведение журнала / восстановление, горячее резервирование и k-безопасность, трехуровневая архитектура для разработки приложений.

12. Сжатие данных в In-Memory базе данных: кодирование длин серий (Run-length encoding), кластерное кодирование (Cluster Encoding), дельта-кодирование (Delta encoding).

13. NoSQL: идея NoSQL, теорема CAP, размещение данных.

14. Параллелизм на уровне аппаратного обеспечения, параллелизм на уровне программного обеспечения: закон Амдала.

Примеры задач для самостоятельного решения.

1. Дана таблица R с атрибутом City размером 50 Byte.

Количество записей в таблице – 14 000 000.

Количество уникальных значений атрибута City – 3000.

Вычислить объем данных в случае:

– построчного хранения;

– поколоночного хранения со словарным кодированием;

– поколоночного хранения со словарным кодированием и со сжатием вектора атрибута с помощью кодирования длин серий (Run Length Encoding).

2. Таблица, содержащая информацию о запасах продукта, имеет следующие атрибуты:

Warehouse (4 byte); Product Id (4 byte); Product Name Short (20 byte); Product Name Long (40 byte); Self Production (1 byte); Production Plant (4 byte); Product group (4 byte); Sector (4 byte); Stock Volume (8 byte); Unit of Measure (3 byte); Price (8 byte); Currency (3 byte); Total Stock Value (8 byte); Stock Currency (3 byte).

Размер строки кэша равен 64 байта. Скорость сканирования ядра процессора составляет 4 МБ в миллисекунду.

Подсчитать время реконструкции кортежа при построчном и поколоночном хранении.

3. Каково максимальное теоретическое ускорение программы:

– состоящей из 50% последовательной части и

– выполняемой четырьмя отдельными процессорами?

4. Сколько бит минимально необходимо для представления 1140 различных значений в векторе атрибута при использовании словарного кодирования?

6. Перечень основной и дополнительной учебной литературы:

А. Основная литература

1. Hasso Plattner. A Course in In-Memory Data Management: The Inner Mechanics of In-Memory Databases. Second Edition. Springer-Verlag Berlin Heidelberg, 2014.

2. Jeffrey Word. SAP HANA Essentials. 5th Edition. Epistemy Press, 2013.

3. Arvind Narayanan, Joseph Bonneau, Edward Felten, Andrew Miller, Steven Goldfeder. Bitcoin and Cryptocurrency Technologies: A Comprehensive Introduction. Princeton University Press, 2016.

Б. Дополнительная литература

1. Hasso Plattner, Bernd Leukert. The In-Memory Revolution: How SAP HANA Enables Business of the Future. Springer International Publishing, 2016.

2. Guy Harrison. Next Generation Databases: NoSQL, NewSQL and Big Data. Apress, 2015.

# 7. Перечень ресурсов информационно-коммуникационной сети «Интернет»:

**Программа утверждена на заседании кафедры теоретической информатики**

**Протокол №**