VCF file storage

Создание интерфейса к базе данных HBase для организации хранения, обработки и предоставления доступа к VCF-файлам

Выполнил: Гайдай И.

Руководитель: Михеев М. (Biodatomics)

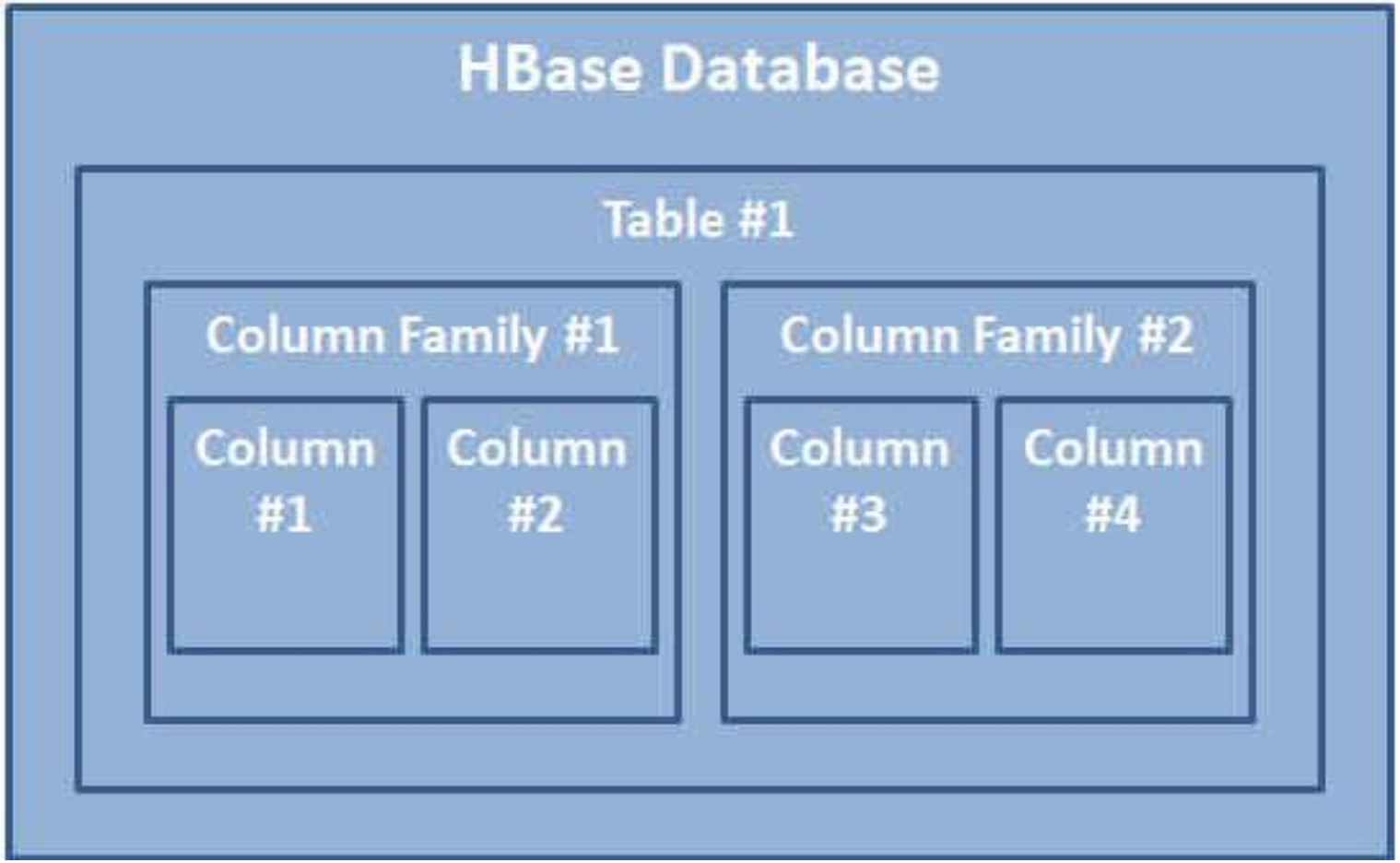
Что планировалось:

Разработать интерфейс к базе данных HBase, позволяющий:

- Производить разбор поступающих VCF-файлов;
- Сохранять VCF-файлы, относящиеся к одному референсу в одной таблице, что позволит хранить файлы более компактно за счет устранения дублирующейся информации, а так же увеличить эффективность обработки за счет представления информации в табличном, а не текстовом виде;
- Получать информацию о сохраненных в базе образцах и формировать текстовое представление VCF-файлов для указаннх column families.

Немного об HBase…

- HBase — нереляционная распределённая база данных.



Что получилось:

- Разработан интерфейс, обладающий описанным выше функционалом;
- Для каждого референса создается отдельная таблица, в которой основным ключом является идентификатор мутации (хромосома, позиция и последовательность нуклеотидов, отличающаяся от референса);
- При добавлении очередного VCF-файла к существующему референсу, список мутаций, которые он описывает добавляется в новую column family;
- Для каждой таблицы можно запросить список VCF-файлов, которые были в ней сохранены (имя файла сохраняется под отдельным ключом) и восстановить VCF-файлы, соответствующие указанным column families.

Результаты:

- Достигнуты поставленные задачи;
- Получен опыт программирования на языке Java, знакомство с фреймворком Maven;
- Получен опыт работы с HBase, а также представление об основных принципах ее работы;
- Знакомство с форматом VCF-файлов.

Дальнейшие шаги:

- Планируется добавить возможность генерировать единый VCF-файл для множества выбранных column families. Данная задача осложняется возможной противоречивостью в описании одних и тех же мутаций в разных VCF-файлах.

Спасибо за внимание!