Группировка больших объемов информации по вычисляемым признакам с помощью реляционных баз данных

Пенских
Юрий Владимирович

Институт солнечно-земной физики СО РАН

ФАЙЛЫ

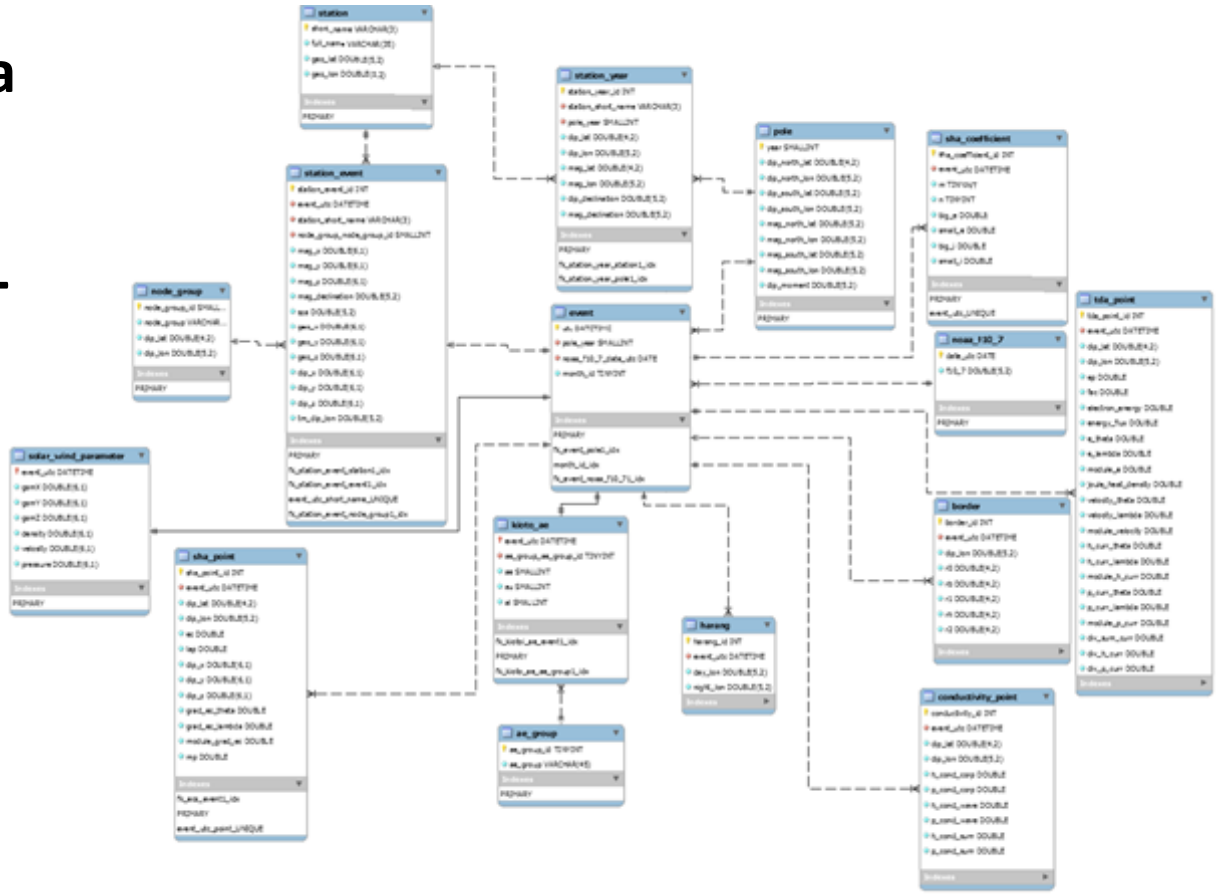
Время	Станция	X	Y	Z
2001-08-27 00:00:00	SIT	7.2	-13.5	2.8
2001-08-27 00:00:00	**PET**	-28.7	15.1	3.1
2001-08-27 00:00:00	**SIT**	7.2	-13.5	2.8

Станция	Полное название	Широта	Долгота
PTK	Paratunka	52.97	158.25
SIT	Sitka	57.07	224.67

ТЕОРИЯ БАЗ ДАННЫХ

ЭДГАР КОДД
Премия Тьюринга (1981)

КРИСТОФЕР ДЕЙТ



Теория множеств
+
Теория предикатов

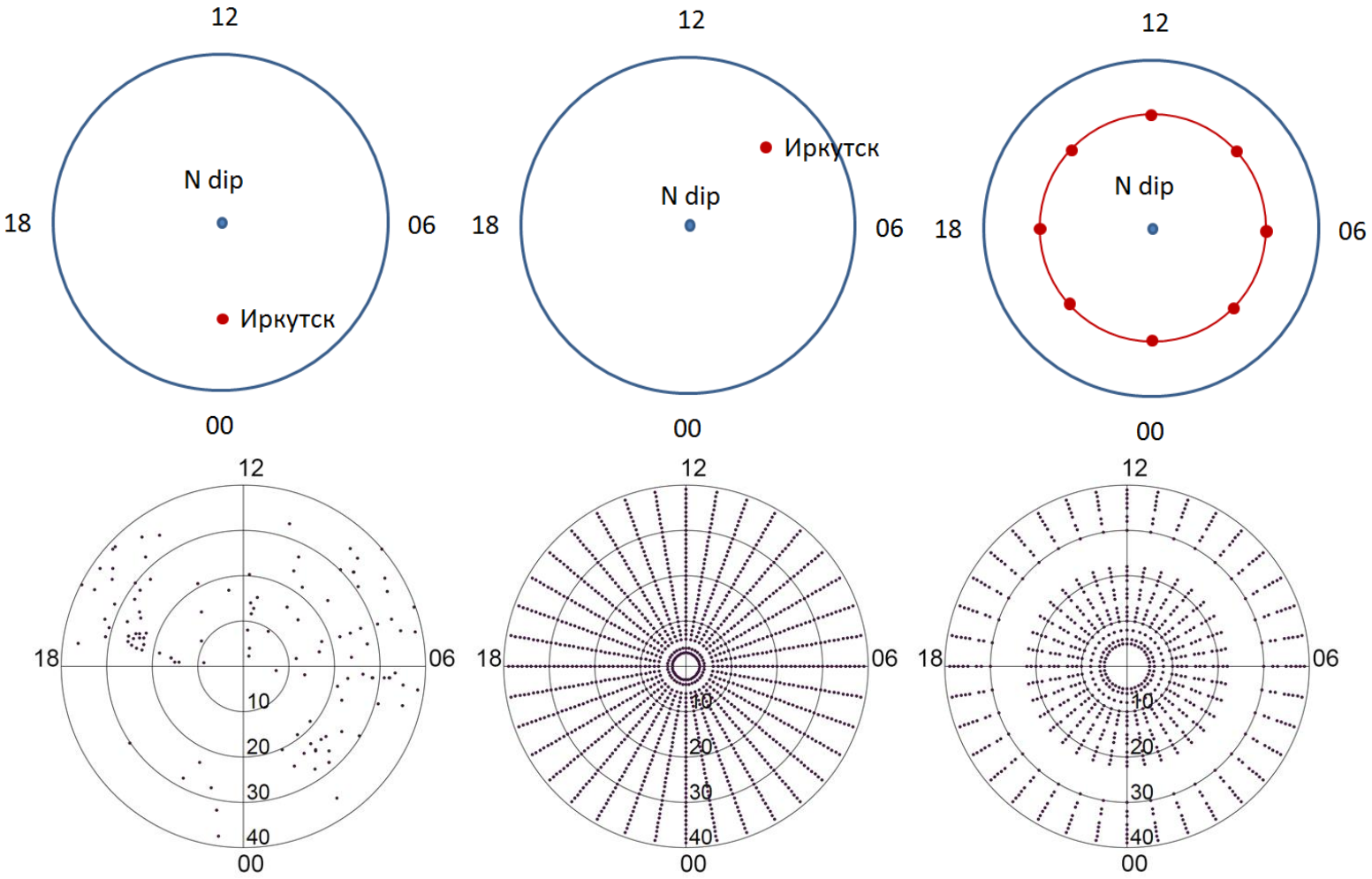
ДАННЫЕ

Существует сеть наземных 288 магнитометров (2008-2009)

- 250 суток минутных данных вариаций геомагнитного поля;
- каждую минуту работает 215 магнитометров;
- индексы геомагнитной активности;
- координаты дипольных полюсов по эпохам;
- и др.

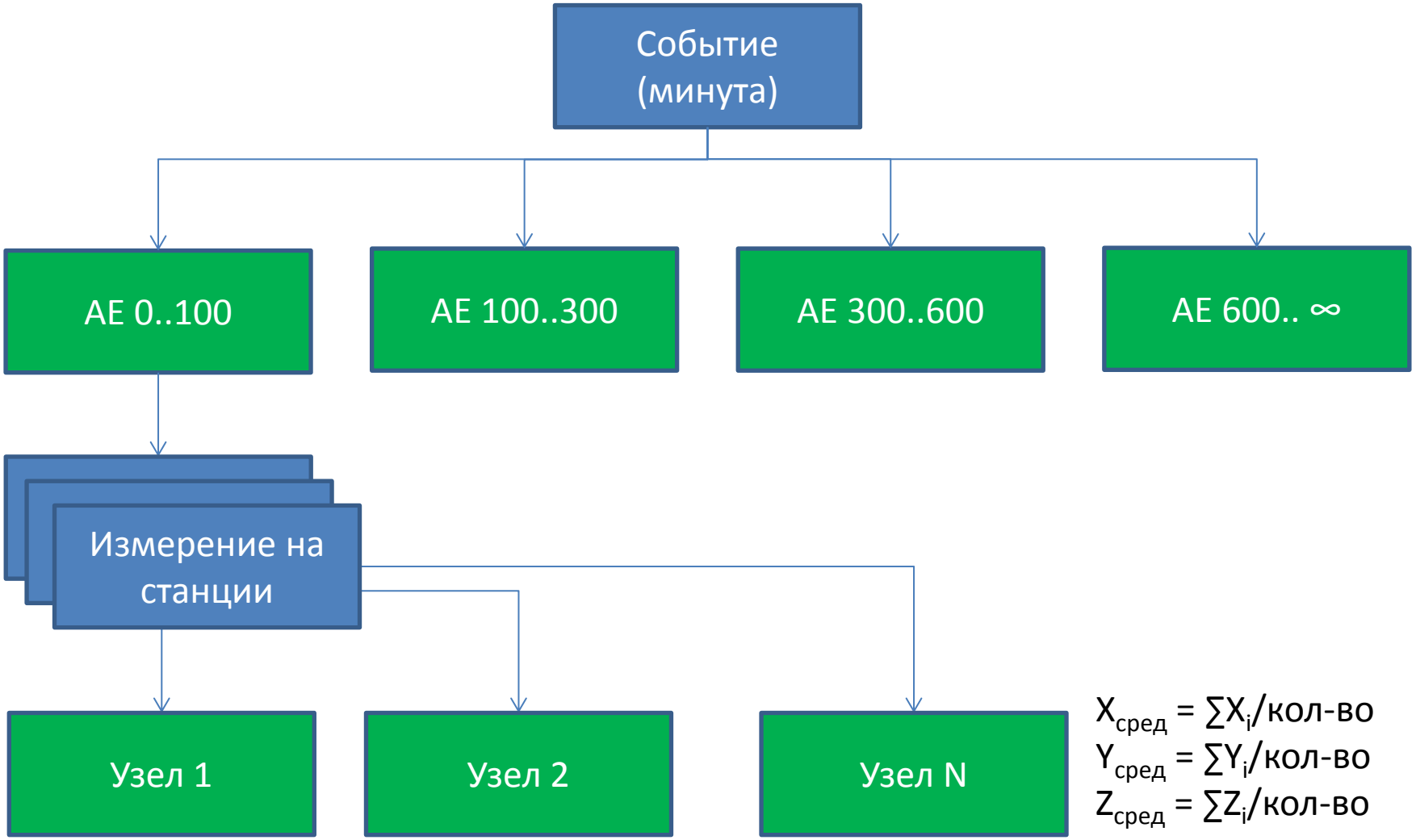
С.К.: дипольная широта (Ф),
местное геомагнитное время(MLT)

ЗАДАЧА



Получаем относительно однородную сеть за сутки

КЛАССИФИКАЦИЯ

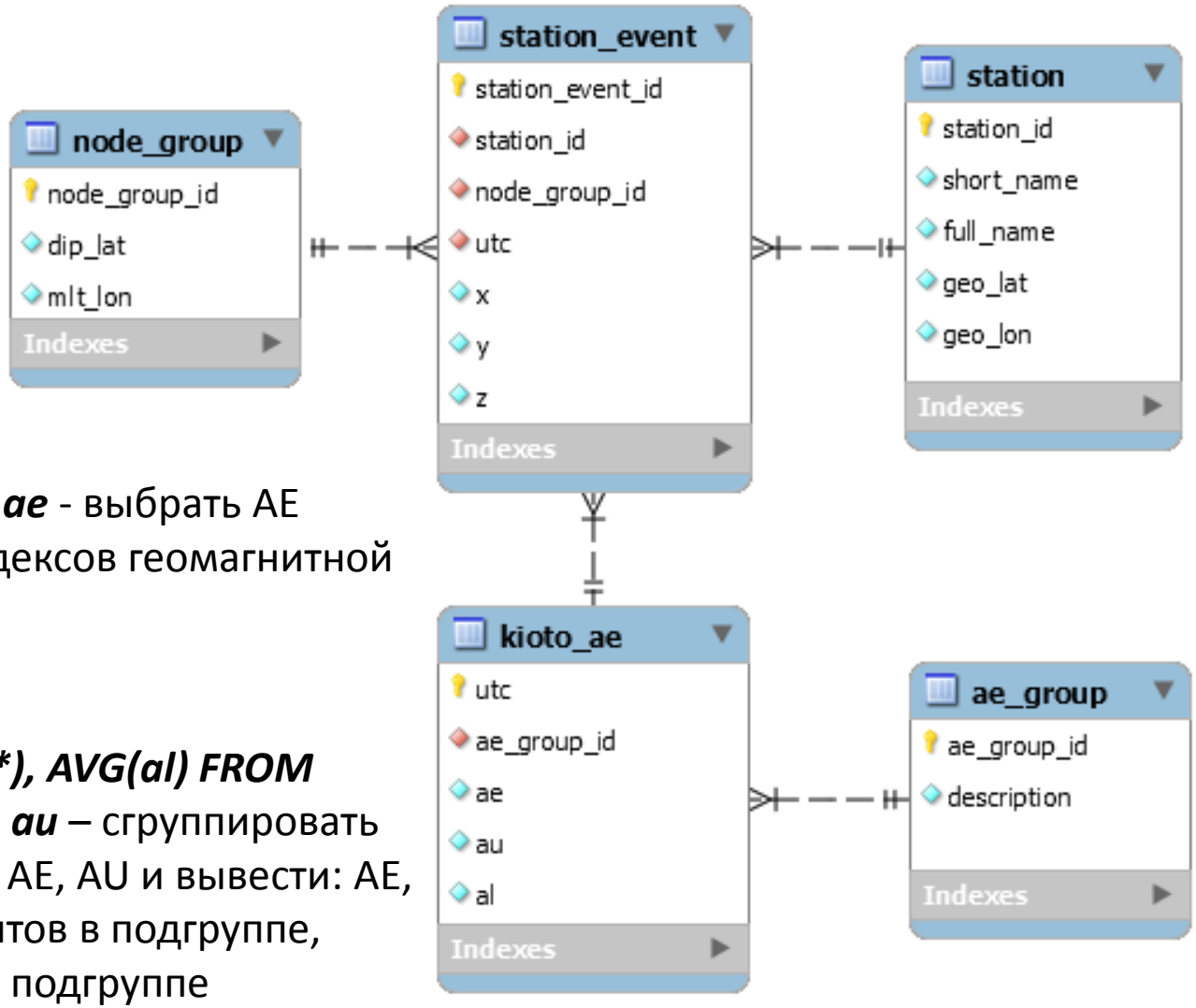


$X_{сред} = \sum X_i/кол\text{-}во$

$Y_{сред} = \sum Y_i/кол\text{-}во$

$Z_{сред} = \sum Z_i/кол\text{-}во$

ГРУППИРОВКА В БД



SELECT ae FROM kioto_ae - выбрать AE индекс из таблицы индексов геомагнитной активности

SELECT ae, au, COUNT(*), AVG(al) FROM kioto_ae GROUP BY ae, au – сгруппировать данные из таблицы по AE, AU и вывести: AE, AU, количество элементов в подгруппе, среднее значение AL в подгруппе

СРАВНЕНИЕ РЕАЛИЗАЦИЙ

Загрузка данных
Расчёт классифицирующий признаков
Группировка данных (программно)

сутки данных – 10 минут
250 суток данных – 1,7 суток

Загрузка данных
Расчёт классифицирующий признаков

сутки данных – 0.7 минуты
250 суток данных – 3 часа

группировка данных (СУБД) – 2 минуты

ИТОГ

1. Задача группировки больших объемов информации по вычисляемым признакам существенно упрощается, если предварительно рассчитать классифицирующие признаки и использовать штатные средства СУБД
2. Проведен сравнительный анализ двух решений задачи (программной и с помощью БД) на примере 250 суток минутных вариаций геомагнитного поля по двум признакам.
3. Показана эффективность реализации с помощью БД

Спасибо
за внимание!