СЛЁРМ Обучение для инженеров и технических лидеров

slurm.io

Дата-инженер

Оценка качества данных

Ася Гайламазян

Оценка качества данных


Statistical biases


Lack of data lineage


Software bugs


Noise


Abnormalities


Information Security


Untrustworthy data sources


Falsification


Uncertainty and ambiguity of data


Duplication of data


Out of date and obsolete data


Human error

Точность, Полнота, Согласованность

Table 2.1. Standard data input tools

Text input box	Pull down list	List	Radio box	Check box
Feb	Feb	Feb Mar Apr May Jun Jul Aug	1 2 3 4	1 2 3 4

Marking Required Fields in Forms

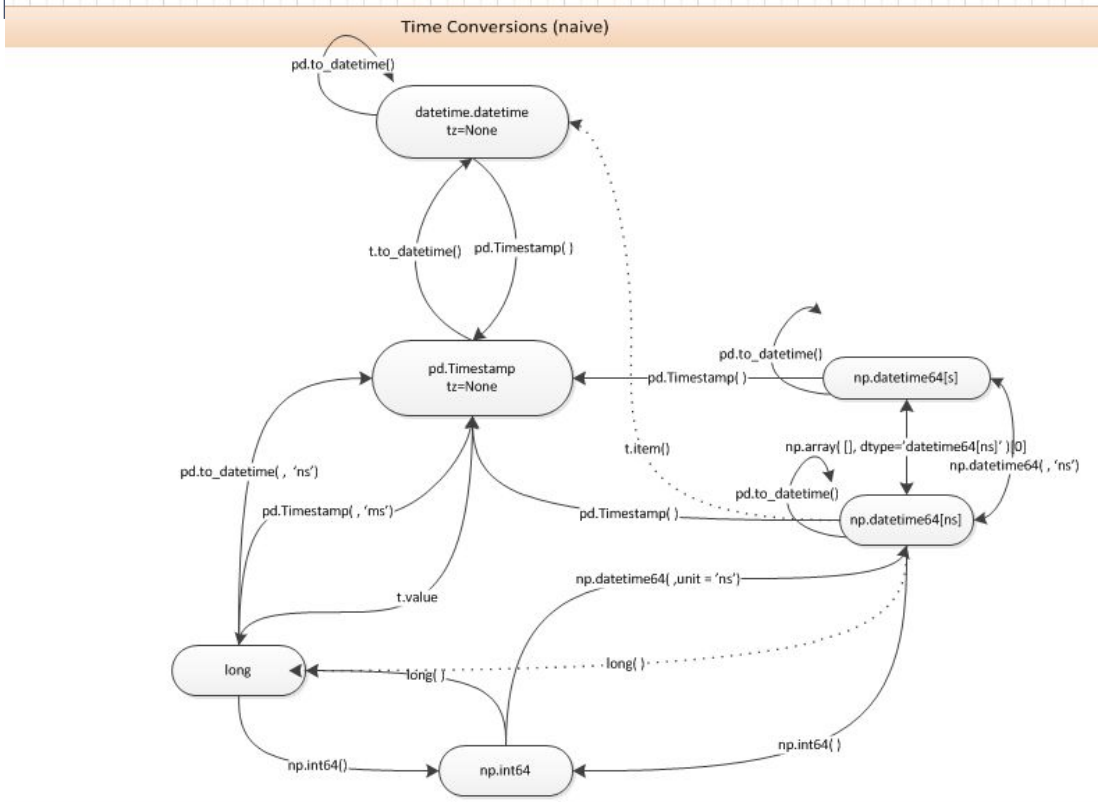
* Name *Required*

* Email Address *Required*

nngroup.com

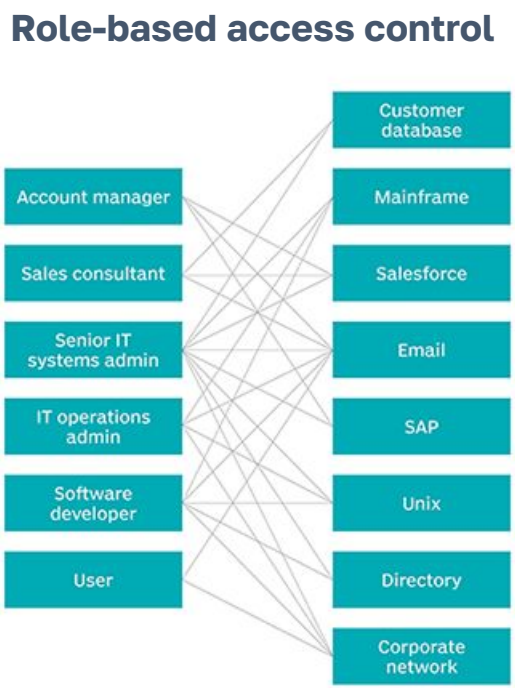
NN/g

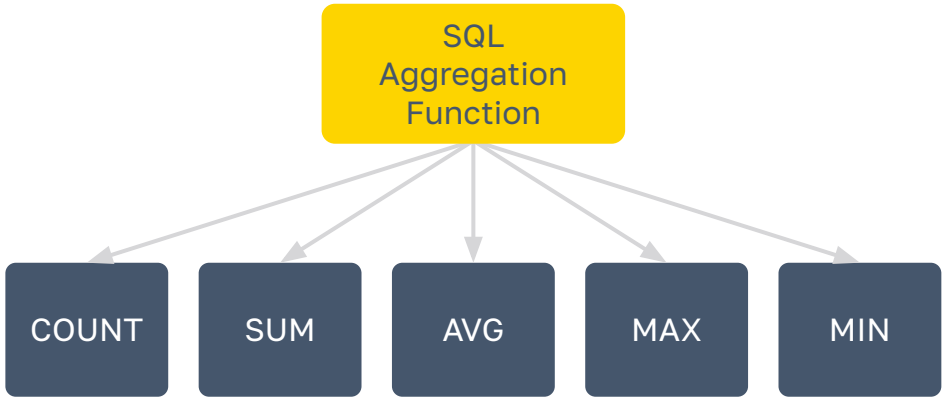
Своевременность, Актуальность, Валидность



Целостность, Дублирование, Уникальность

Четкость, Надежность, Доступность





Возможность аудита, документация, этические соображения

Три простых шага аудита данных

1. Привлеките заинтересованные стороны
Данные, важные для опыта клиентов, вероятно, хранятся и используются в разных отделах. Найдите и привлеките ключевые заинтересованные стороны, которые могут рассказать о процессах сбора, хранения и использования данных.

2. Составьте карту расположения ваших данных
Найдите все места, где хранятся данные, важные для обслуживания клиентов, и наметьте информационную архитектуру этих данных, включая то, как они хранятся и кто имеет доступ.

3. Оцените точность, широту и последовательность
Углубитесь и оцените качество своих данных, используя принципы точности, широты и последовательности, а затем проведите мозговой штурм по решению любых проблем, которые вы обнаружите в ходе аудита.

kafka

Microsoft SQL Server

amazon Kinesis

JSON, CSV, TXT...

БРОНЗОВЫЙ

Необработанные данные и история

Складывание данных после получения
Источники

СЕРЕБРЯНЫЙ

Фильтрация, очищение и дополнение

Обработка недостающих данных
Стандартизирование чистых полей
Демультиплексирование вложенных объектов
Удобное название полей
Обогащение

ЗОЛОТОЙ

Уровень бизнеса Агрегация

Бизнес модель
Агрегация для измерений
Удобные для бизнеса названия полей

Data Quality

Stream analytics

BI Reporting

Data Science & ML