HUAWEI

Day10 实时AI:StreamingML和实时可视化实践

机器学习应用广泛，在线机器学习是空白

机器学习算法：回归、分类、聚类

监督学习、非监督学习、半监督学习

离线训练模型，训练一次模型，使用模型做预测。

但，

模型的训练集在持续增加，而模型一直不变，模型越旧，预测结果越难匹配最新数据

HUAWEI

什么是在线机器学习StreamingML

定义：模型训练和预测同时进行，模型更新频率为秒级或分钟级

描述：不同于离线训练，模型在窗口时间内做迭代训练，增量更新模型

擅长：在实时场景，离线模型比较之后，实时模型能最大程度适应最新的流数据

HUAWEI

StreamingML的优势

- 模型秒级更新
- 支持监督、非监督、半监督模型
- StreamSQL实现ML算法

实时流计算服务Console页面


```
/**
  * 该示例为流分析场景通用模板。数据的输入源和输出通道均由DIS服务提供，需先开通DIS服务并创
  * >>>>>>>>>>请务必确保您的账户下已在数据接入服务（DIS）里创建了您配置的通道<<<<<<<<<<
  *
  * >>>>>样例输入<<<<<
  *  流名：car_infos(car_id,car_owner,car_brand,car_price):
  *  1,lilei,bmw320i,28
  *  2,hanmeimei,audia4,27
  * >>>>>样例输出<<<<<
  *  流名：audi_cheaper_than_30w(car_id,car_owner,car_brand,car_price):
  *  2,hanmeimei,audia4,27
  **/

/** 创建输入流，从DIS的csinput通道获取数据。
  *
  * 根据实际情况修改以下选项：
  * channel：数据所在通道名
  * partition_count：该通道分区数
  * encode：数据编码方式，可以是csv或json
  * field_delimiter：当编码格式为csv时，属性之间的分隔符
  **/
CREATE SOURCE STREAM car_infos (
  car_id STRING,
  car_owner STRING,
  car_brand STRING,
  car_price INT
)
WITH (
  type = "dis",
  region = "cn-north-1",
  channel = "csinput",
  partition_count = "1",
```

最强StreamSQL，智能流计算

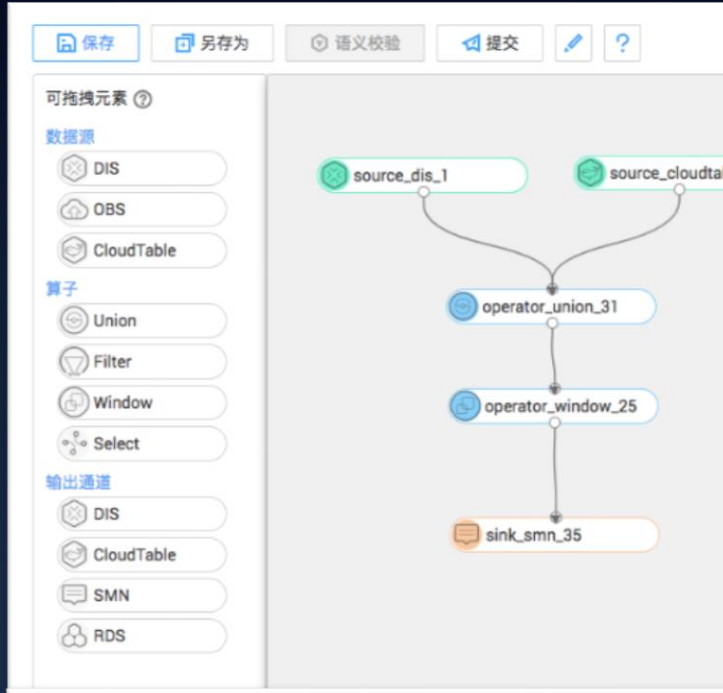
Flink CEP SQL：规则引擎

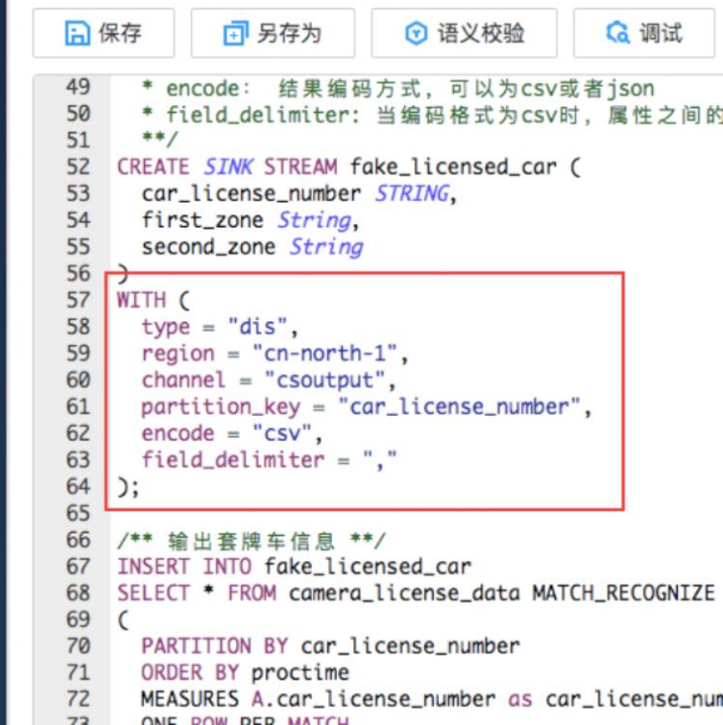
Time GeoSpatial：时间+地理位置分析

Job as a Service：Sink流数据订阅推送

StreamingML：20+流式ML算法

独享集群：物理隔离，按需付费，弹性伸缩





应用场景

车联网&物联网

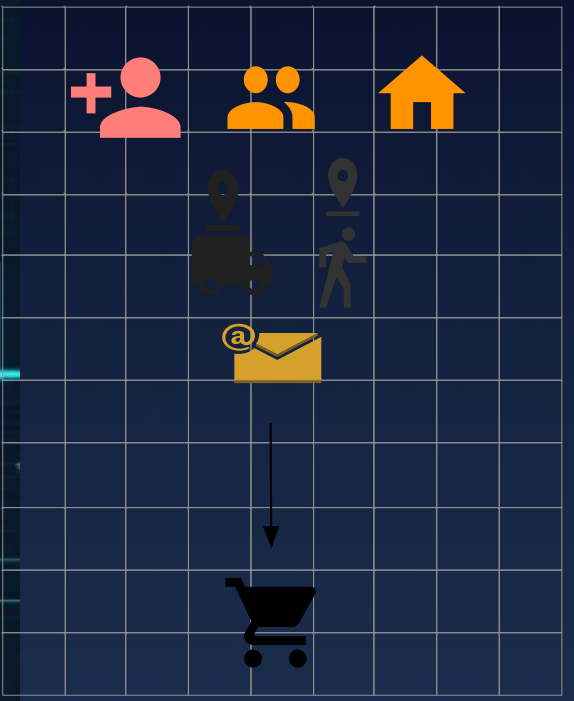
电子围栏
偏航检测
超速检测
频繁轨迹
危险驾驶行为分析
…

金融&交易所

实时风控
拉抬打压
异常监控
K线计算
实时指标

交通智能体

信号优化
车辆管理
故障检测
在线机器学习



通用场景

ETL数据清洗
特征工程
在线机器学习
实时推荐
日志分析

HUAWEI

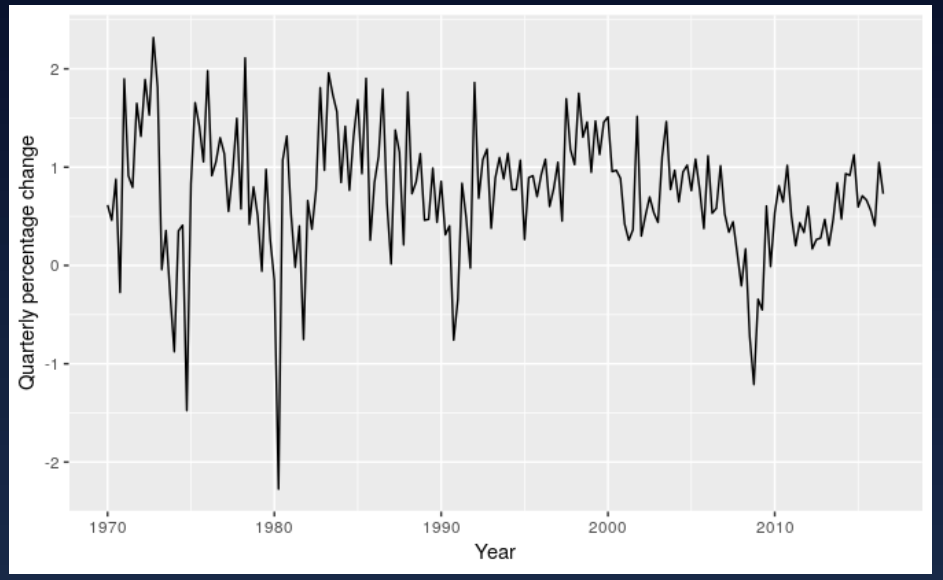
常见的StreamingML算法

1. 时间序列预测
2. 异常检测
3. 实时聚类
4. 统计分析

HUAWEI

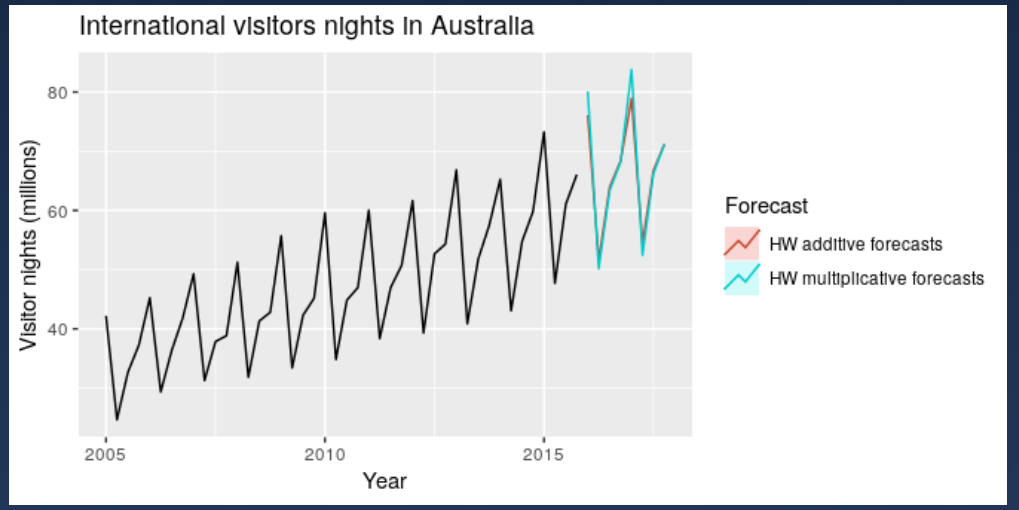
StreamingML - 时间序列预测

ARIMA
非季节性序列预测




```
SELECT
ARIMA_PRED(b) OVER (ORDER BY rowtime ROWS  BETWEEN 5
PRECEDING AND CURRENT ROW) AS arima
FROM MyTable
```

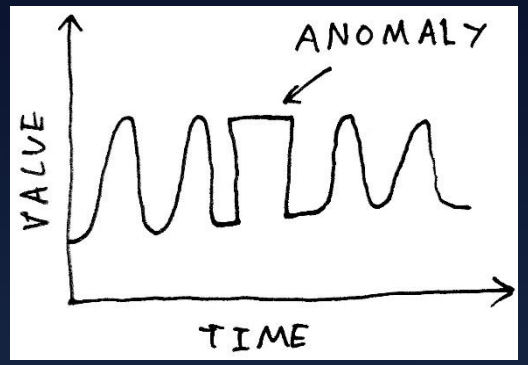
HOLT-WINTERS
季节性序列预测

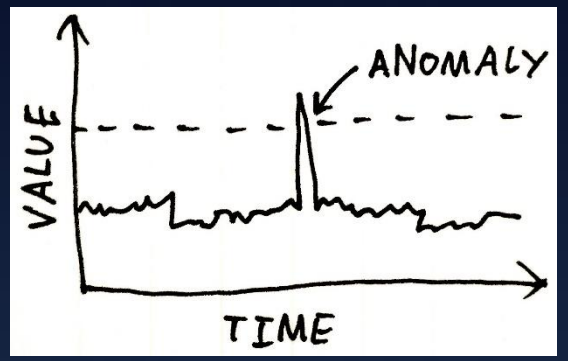


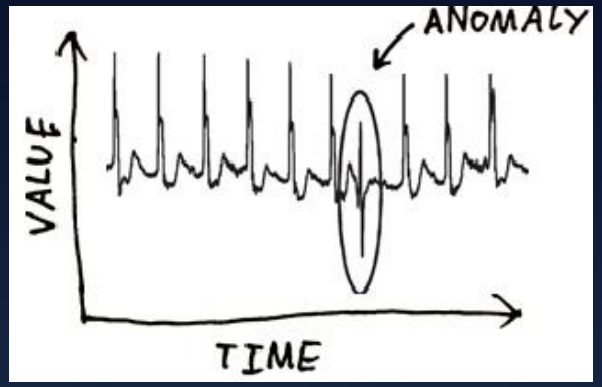

```
SELECT
HOLT_WINTERS(b, 5, 2) OVER (ORDER BY rowtime ROWS  BETWEEN
5 PRECEDING AND CURRENT ROW) AS hw2,
FROM MyTable
```

HUAWEI

StreamingML – 异常检测

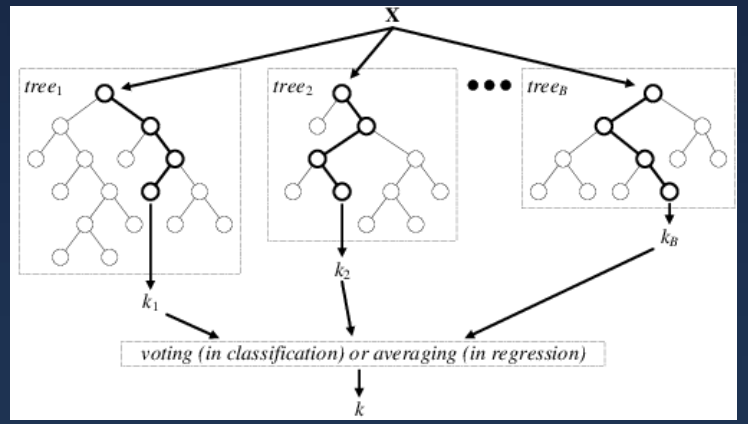






流式随机森林

入侵检测
金融欺诈检测
数据监控
医疗诊断



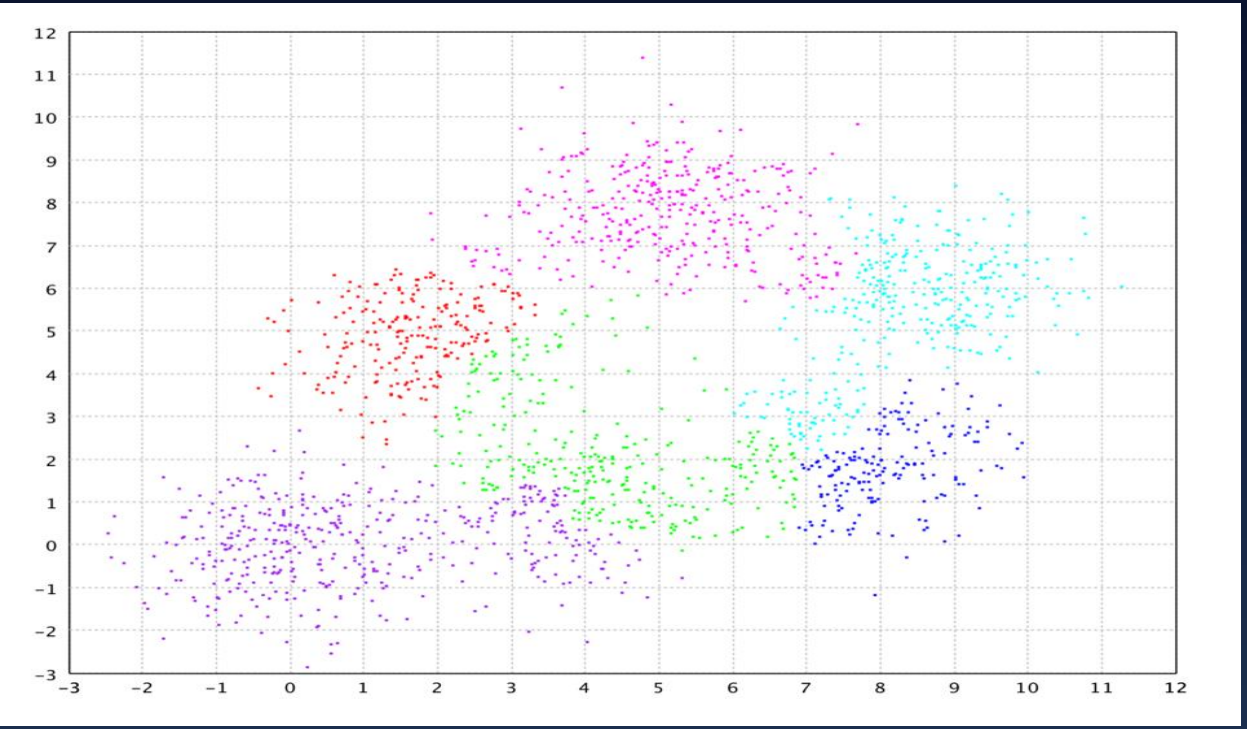

```
SELECT SRF_UNSUP(ARRAY[c]) OVER (
       ORDER BY rowtime RANGE BETWEEN INTERVAL '3' SECOND PRECEDING AND CURRENT ROW) AS unsup
FROM mytable
```

HUAWEI

StreamingML – 实时聚类

Adaptive Clustering

无需设定聚类数目
低延时

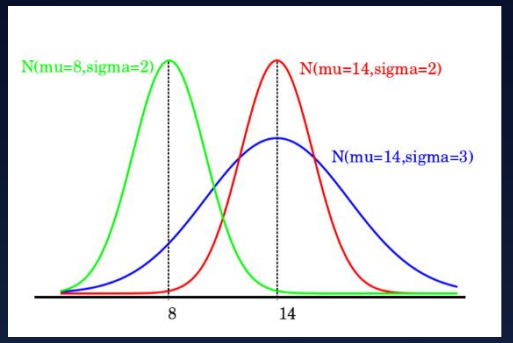


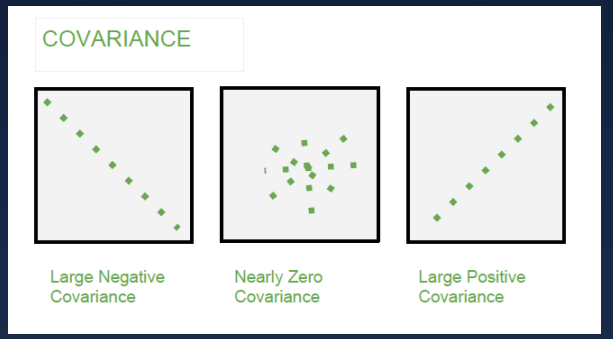

```
SELECT
CENTROID(ARRAY[c,e], 1.0) OVER (ORDER BY proctime RANGE UNBOUNDED PRECEDING) AS centroid,
CLUSTER_CENTROIDS(ARRAY[c,e], 1.0) OVER (ORDER BY proctime RANGE UNBOUNDED PRECEDING) AS centroids
FROM MyTable
```

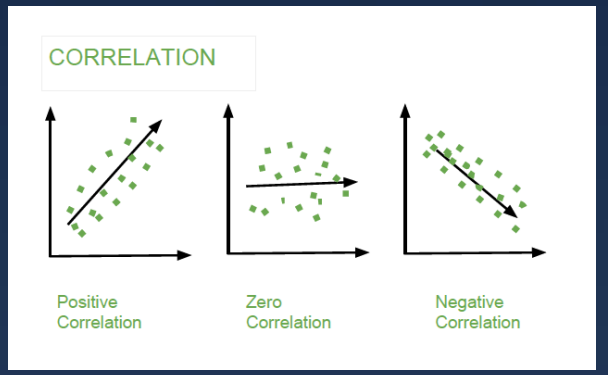
HUAWEI

StreamingML – 统计分析

- `SUM(field_name)`
- `COUNT(field_name)`
- `AVG(field_name)`
- `MIN(field_name)`
- `MAX(field_name)`
- `STD(field_name)`
- `VAR(field_name)`
- `SUM_MULTI_ARG(field_name1, field_name2)`
- `COVAR(field_name1, field_name2)`
- `CORR_COEF(field_name, field_name)`
- `ARGMIN(field_name, UDF(field_name))`
- `ARGMAX(field_name, UDF(field_name))`







HUAWEI

实时流计算服务CS

Thank You.

https://www.huaweicloud.com/product/cs.html