

ArchData

技术峰会北京站

主办方：



2017年9月24日北京海淀区丹棱街5号微软亚太研发中心一号楼一层 故宫会议室



AI集成之大数据平台建设

自我介绍

何文斌

2010年北师大毕业

Amazon, SDE, FC Capacity Planning

各种创业公司辗转

融数数据大数据平台负责人

议题

1.

大数据平台功能和架构体系

2.

数据实验室建设和AI集成

3.

大数据运维实践

大数据平台建设愿景和目标

1

任意输入，任意处理，
任意输出

2

打通所有数据工作者所有数据
处理流程

3

集成 AI工具和算法


4

使用方便高效

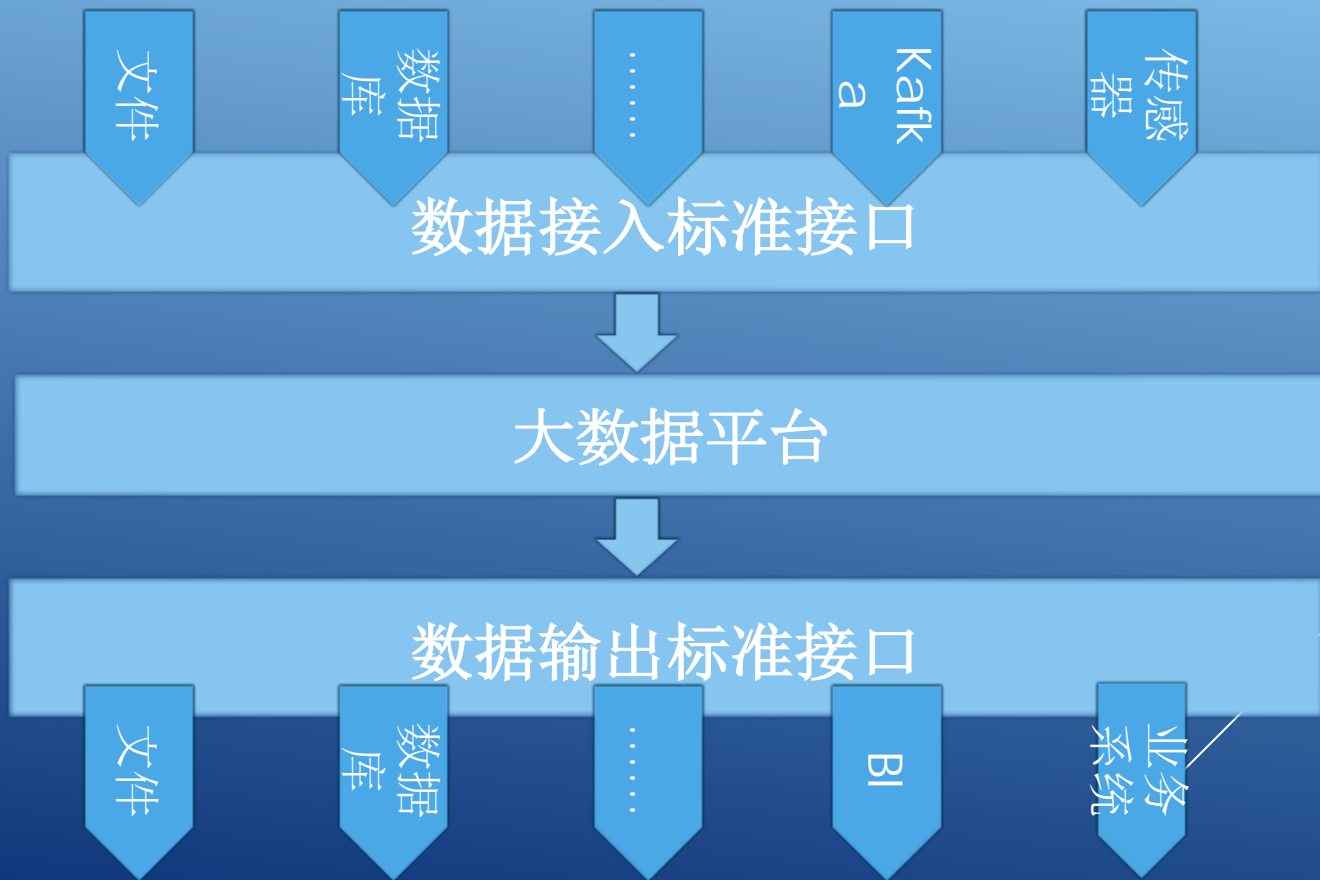
大数据平台功能和架构体系



大数据平台的设计思想

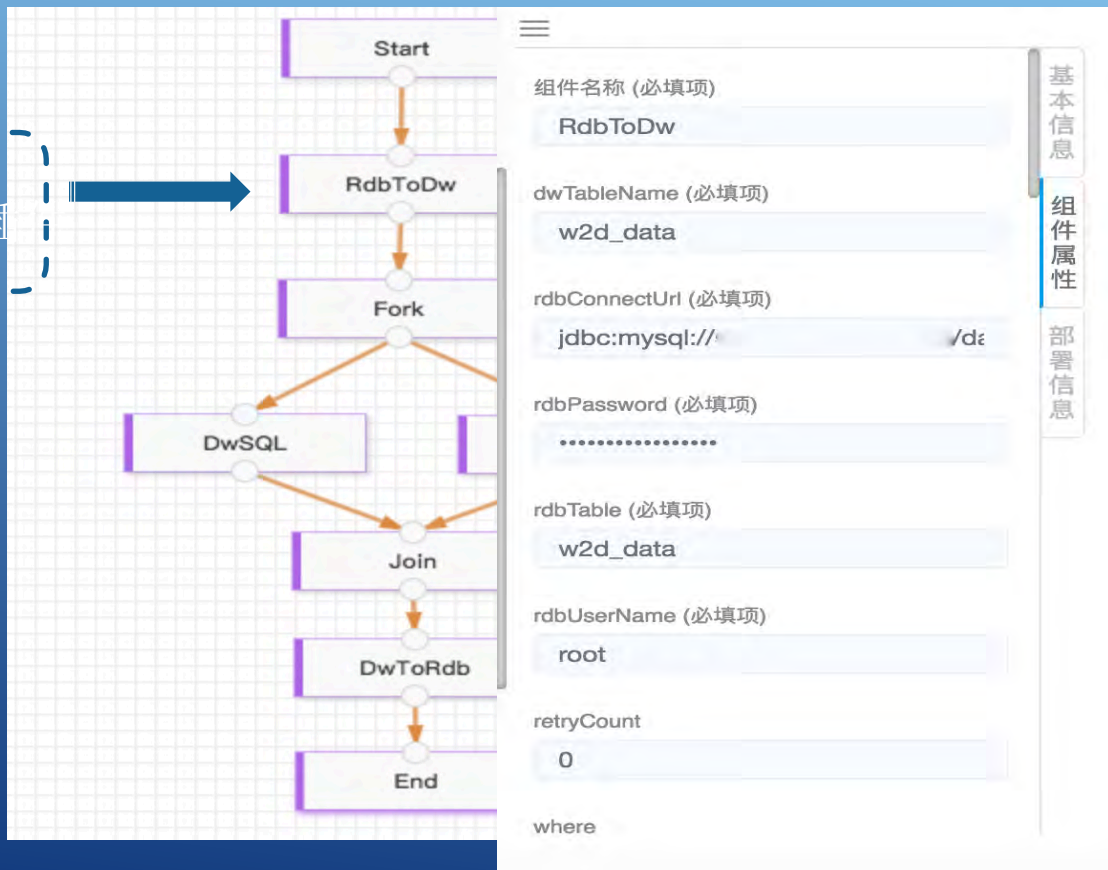
- ◆ 以 Hadoop 为基础
 - ◆ 以开放组件为中心
 - ◆ 以任务为组织
 - ◆ 以血缘为线索
- 

大数据平台之数据采集



大数据平台之任务编排调度

数据处理插



大数据平台组件

```
public class ImportDataMetadata {  
  
    @PaneElement(TxnInitText)  
  
    @Bean  
    public Step importDataStep1(  
        @Value("${dwConnectUrl}") String dwConnectUrl,  
        @Value("${dwDbName}") String dwDbName,  
        @Value("${dwTableName}") String dwTableName,  
        @Value("${dwUserName}") String dwUserName,  
        @Value("${dwLocation}") String dwLocation,  
        @Value("${rdbConnectUrl}") String rdbConnectUrl,  
        @Value("${rdbUserName}") String rdbUserName,  
        @Value("${rdbPassword}") String rdbPassword,  
        @Value("${rdbTable}") String rdbTable,  
        @Value("${where}") String where,  
        @Value("${columnNameMapStr}") String columnNameMapStr,  
        @Value("${partitionMapStr}") String partitionMapStr,  
        @Value("${retryCount}") Integer retryCount,  
        @Value("${dataCoverStrategy}") String dataCoverStrategy,  
        @Value("${importStrategy}") String importStrategy  
    ) {  
        return steps.get("importDataStep1").tasklet(new SimpleTasklet() {  
            @Override  
            public RepeatStatus execute(ChunkContext chunkContext, StepContribution contribution) throws Exception {...}  
  
            private String combineTypeAndLength(String columnName, int precision, int scale) {  
                if ("varchar".equalsIgnoreCase(columnName) || "char".equalsIgnoreCase(columnName)){  
                    return String.format("%s (%d)", columnName, precision);  
                } else if ("decimal".equalsIgnoreCase(columnName)){  
                    return String.format("%s (%d,%d)", columnName, precision, scale);  
                } else {  
                    return columnName;  
                }  
            }  
        }  
    }.retry(retryCount)).build();  
}
```

大数据平台之数据管理

元数据管理

◆ 文件元数据

数据管理 / 数据表管理

t_lgst_comp

字段信息

字段名称	字段类型	字段长度	是否主键	是否外键	是否为空	是否索引
SHOP_GUID	STRING	10	否	否	否	否
LGST_NAME	STRING	10	否	否	否	否
POST_CODE	STRING	10	否	否	否	否
CARD_ID	STRING	10	否	否	否	否
TEL	STRING	10	否	否	否	否

是否支持快照: 否

是否对外开放: 否

大数据平台之数据存储和管理



大数据平台之数据安全



议题

1.

大数据平台功能和架构体系

2.

数据实验室建设和AI集成

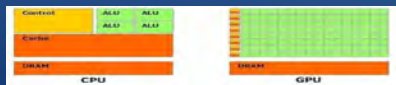
3.

大数据运维实践

AI时代大数据平台建设

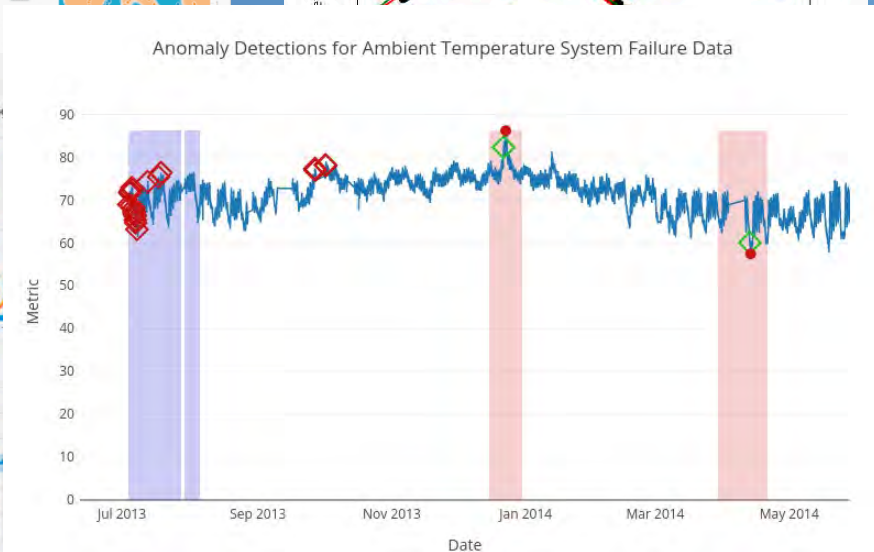
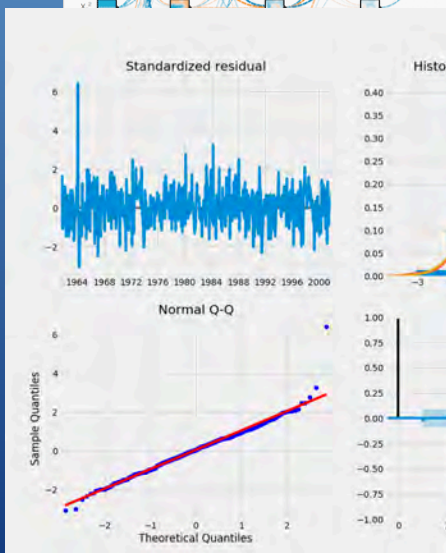
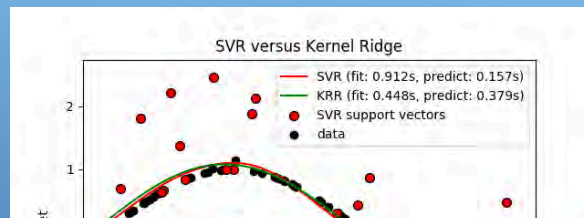
数据实验室建设

PYTHON SQL JAVA



AI集成

数据实验室建设





时间序列数据分析

```
%python
import warni
import itert
import panda
import numpy
import stats
import matplo

plt.style.us
plt.xlim([1,
data = sm.dat
y = data.data

# The 'MS' s
y = y['co2']

# The term b
y = y.fillna
print(y)
```

TensorFlow 深度学习

```
%python
#MNIST 手写数字识别
from tensorflow.examples.tutorials.mnist import input_data
import tensorflow as tf

mnist = input_data.read_data_sets("MNIST_data/",one_hot=True)

x = tf.placeholder("float", [None, 784])

W = tf.Variable(tf.zeros([784,10]))
b = tf.Variable(tf.zeros([10]))

y = tf.nn.softmax(tf.matmul(x,W) + b)

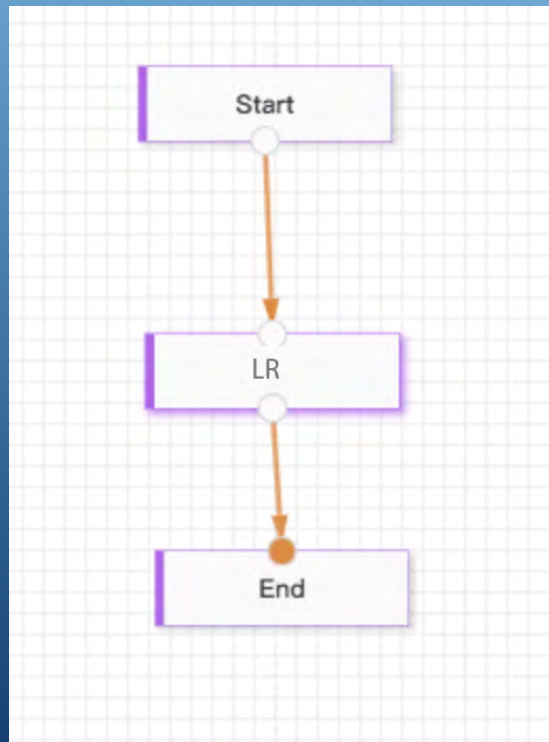
y_ = tf.placeholder("float", [None,10])

cross_entropy = -tf.reduce_sum(y_*tf.log(y))

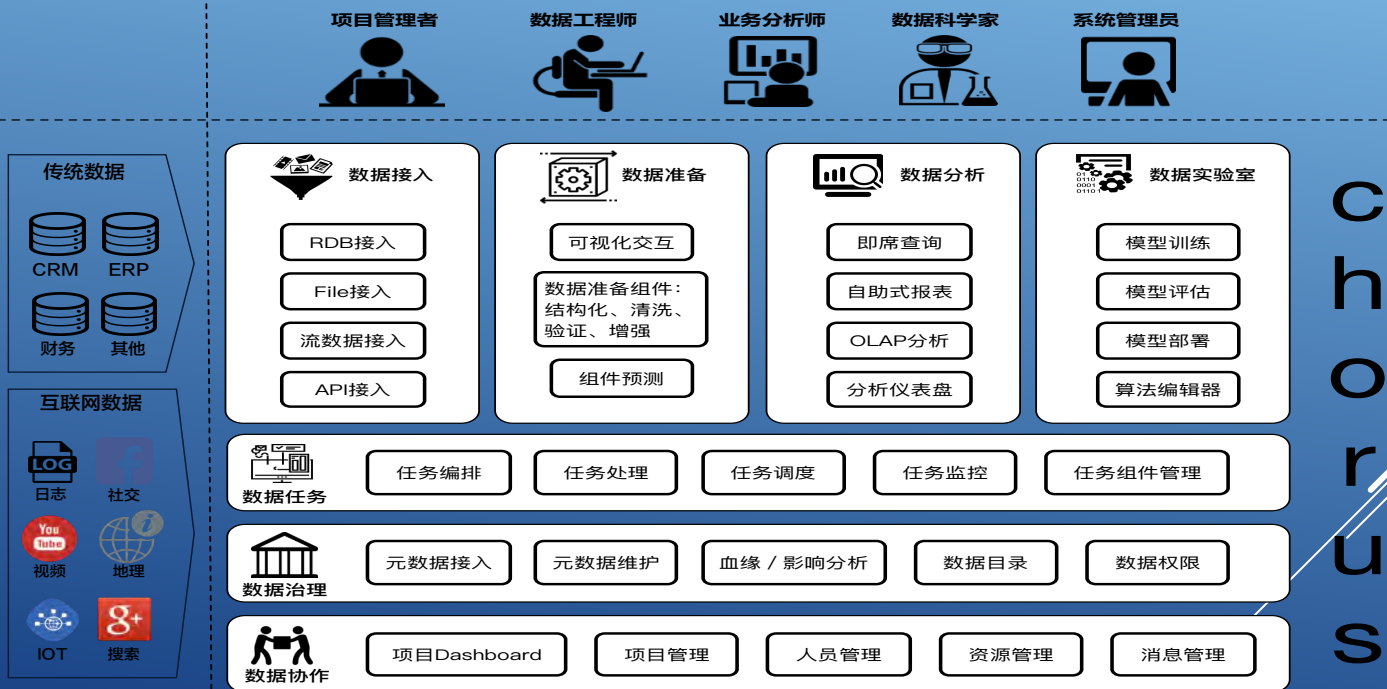
#卷积神经网络模型
def weight_variable(shape):
    initial = tf.truncated_normal(shape, stddev=0.1)
    return tf.Variable(initial)

def bias_variable(shape):
    initial = tf.constant(0.1, shape=shape)
    return tf.Variable(initial)
```

算法组件



大数据平台产品架构



议题

1. 大数据平台功能和架构体系

2. 数据实验室建设和AI集成

3. 大数据运维实践

运维大数据实践

指标管理

业务系统

报警系统

数据输出插件（异常点）

监督学习

无监督学习

数据平台

指标

数据接入处理插件

日志

硬件性能
数据

AP
M

运营

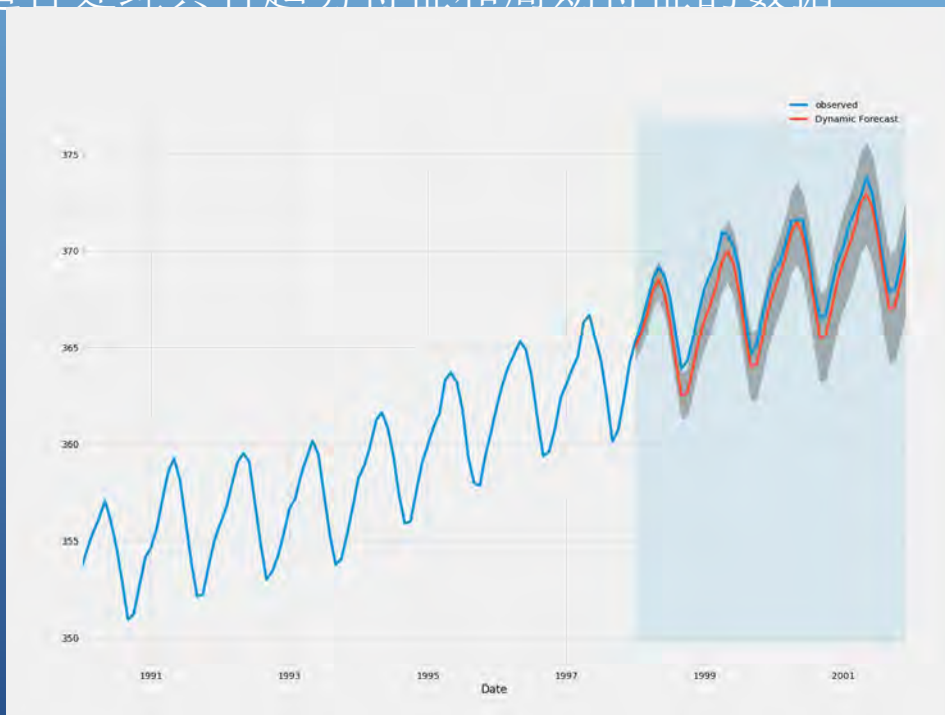
运维大数据实践



运维大数据实践

SARIMA

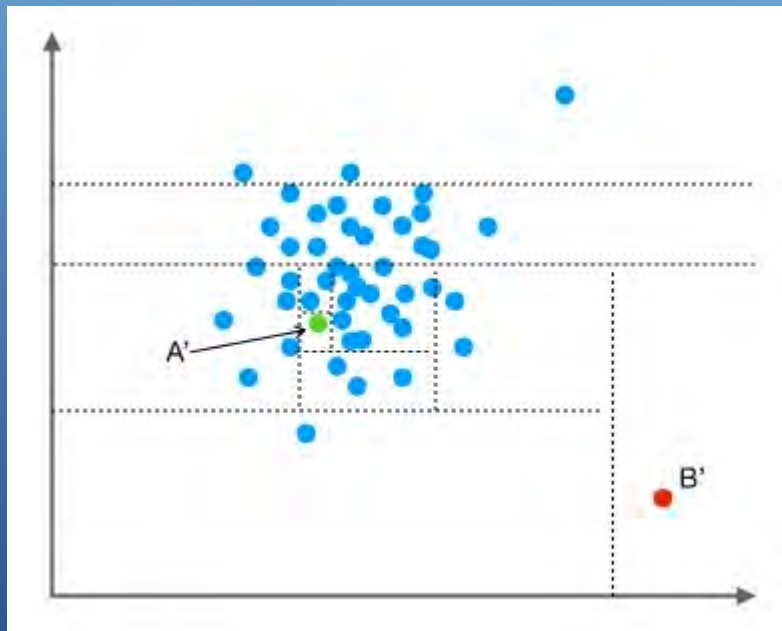
适合处理具有趋势特征和周期特征的数据



运维大数据实践

Isolation Forest

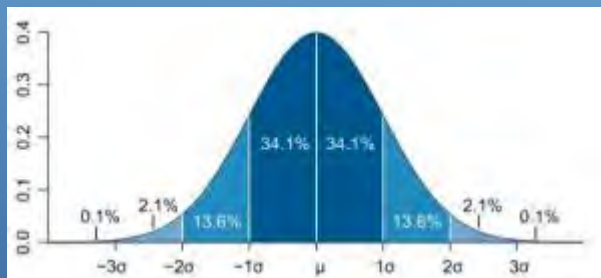
适合多纬度数据



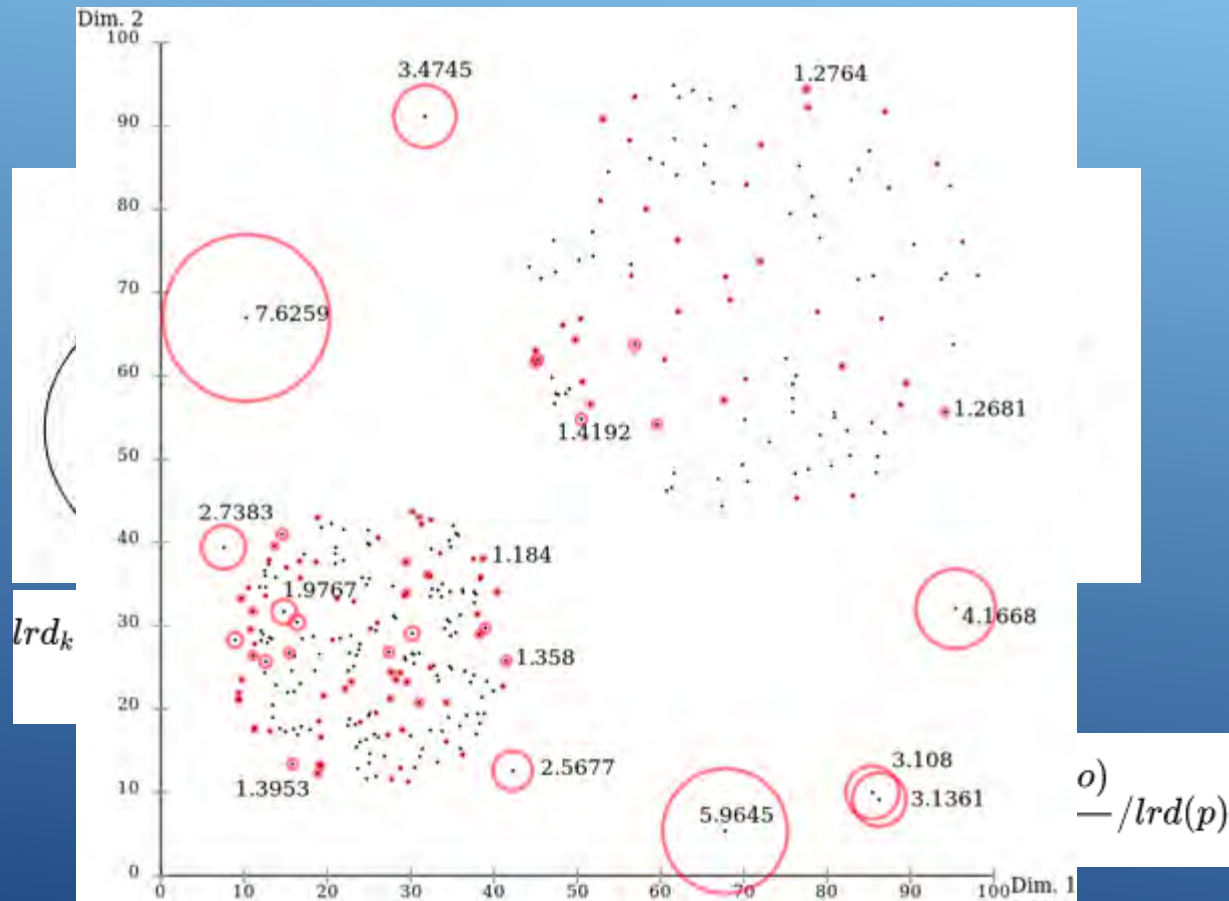
运维大数据实践

Covariance Estimate

适合数据满足高斯分布



运维大数据实践





Thanks!



中生代技术

FRESHMAN TECHNOLOGY



ArchData技术峰会全国巡回

上海9月, 北京9月, 成都10月, 南京10月,
长沙11月, 广州11月

中生代咨询内训

技术架构, 研发管理, 敏捷开发, 大数据
微服务, AI, 机器学习

中生代人才内推

对接研发主管, 内推精准人才