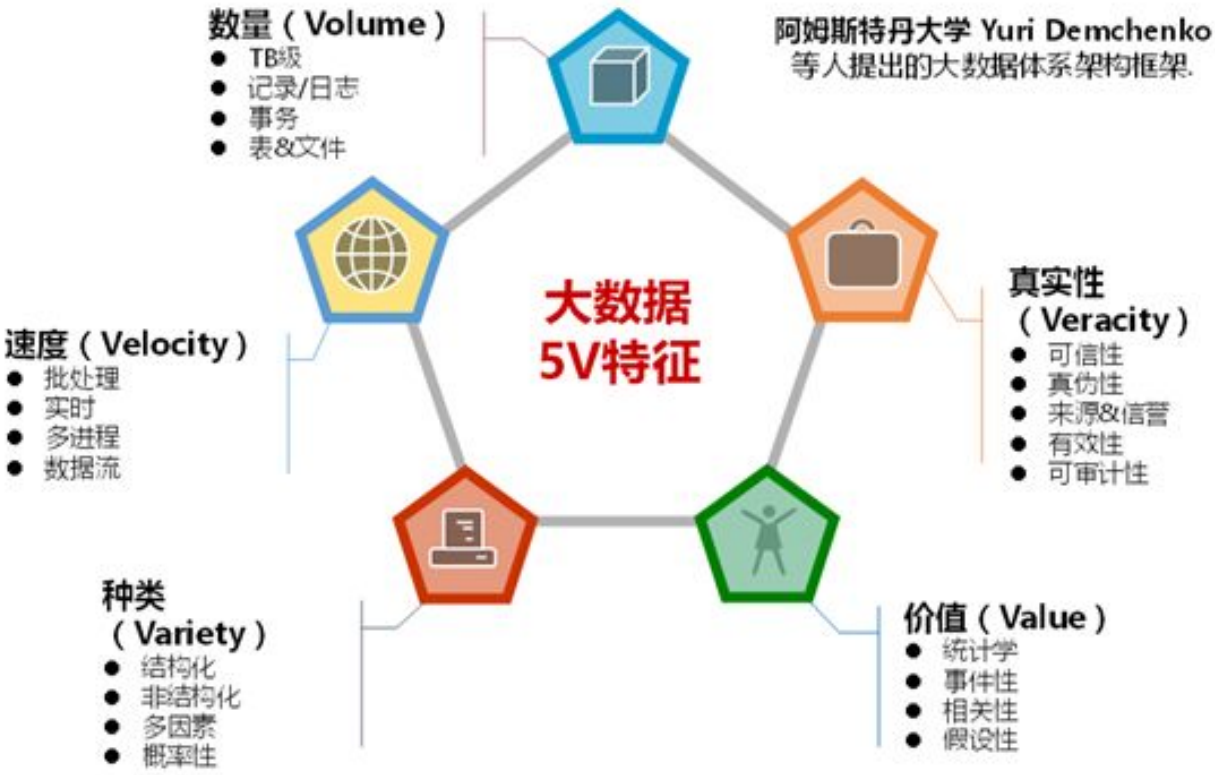
第一节、大数据概述

一、大数据

1. 基本概念

> 巨量数据集合,无法在一定时间范围内用常规软件工具进行捕捉、管理和处理的数据集合



大数据特征

- 数据
- 存储
- 软件
- 处理

> 大数据技术的战略意义不在于掌握庞大的数据信息，而在于对这些含有意义的数据进行专业化处理。

2. 应用场景

- 客户分析

> 分析客户的信息资料、行为和特点，对客户进行细分、预测流失等。

- 营销分析

> - 使用营销模型改进面向客户的应用程序，更好的向客户推荐
> - 分析营销过程，并做出相应调整，优化绩效。

- 社交媒体分析

> 获取不同社交媒体渠道生成的内容做为分析客户情感和舆情监督的基础。

- 网络安全

> 建立分析模型，监测大量网络活动数据和相应的访问行为，以识别可能进行入侵的可疑模式。

- 设备管理

> 收集和分析传感器数据流，包括连续用电、温度、湿度和污染物颗粒等无数潜在变量。以预测设备故障，安排预防性的维护，确保项目正常进行。

- 供应链和渠道分析

> 通过对仓库库存，POS交易和多种渠道的运输进行分析，建立预测分析模型，有效帮助预先补货和管理物流。

- 价格优化

> 通过收集不同种类的数据流，包括竞争对手的价格、不同地域的销售交易数据等，建立分析模型，以达到产品销售的利益最大化。

- 欺诈行为检测

> 通过对上亿条的交易数据进行分析，以识别欺诈行为模式

二、云计算

1. 云的概念

> 按使用量付费的模式，这种模式提供可用的、便捷的、按需的网络访问， 进入可配置的计算资源共享池（资源包括网络，服务器，存储，应用软件，服务），这些资源能够被快速提供，只需投入很少的管理工作，或与服务供应商进行很少的交互。

云

- 按量付费
- 资源共享

2. 云计算应用

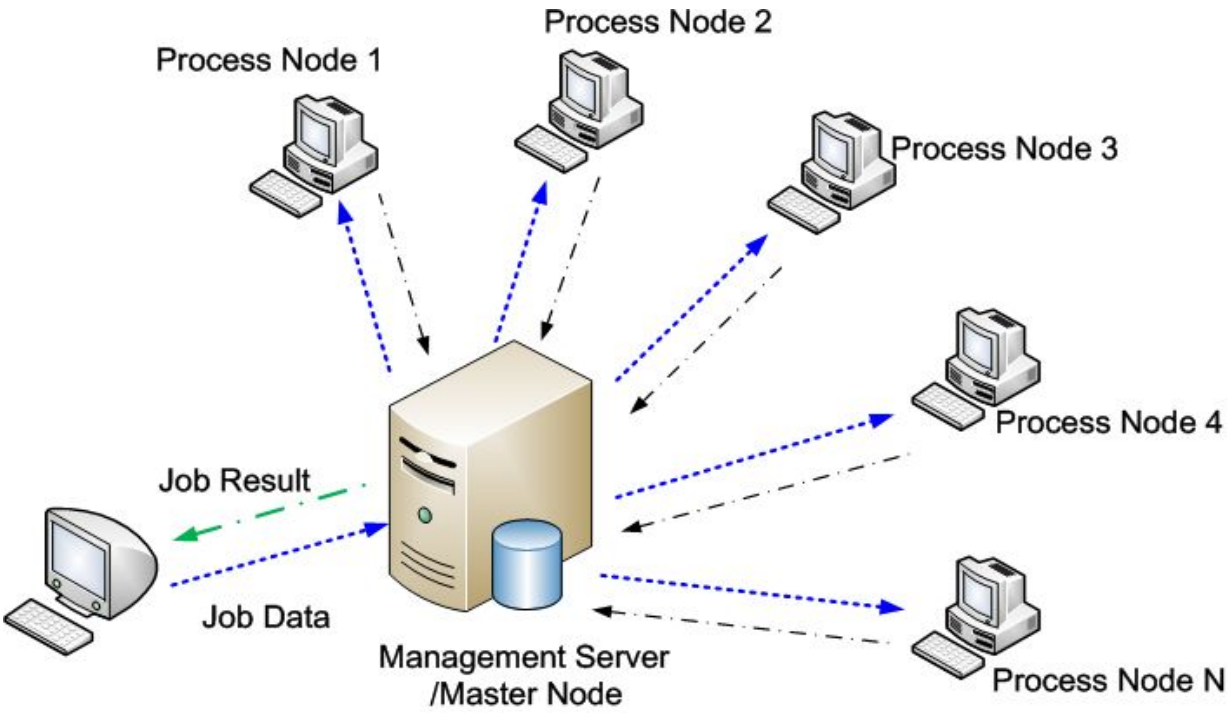
> 云计算（Cloud Computing）是分布式计算（Distributed Computing）、并行计算（Parallel Computing）、效用计算（Utility Computing）、网络存储（Network Storage Technologies）、虚拟化（Virtualization）、负载均衡（Load Balance）、热备份冗余（High Available）等传统计算机和网络技术发展融合的产物。

- 分布式计算
- 并行计算

三、分布式系统

1. 分布式概念

> 分布式系统（distributed system）是建立在网络之上的软件系统。分布式系统具有高度的内聚性和透明性。内聚性是指每一个数据库分布节点高度自治，有本地的数据库管理系统。透明性是指每一个数据库分布节点对用户的应用来说都是透明的，看不出是本地还是远程。



分布式集群

2. 应用领域

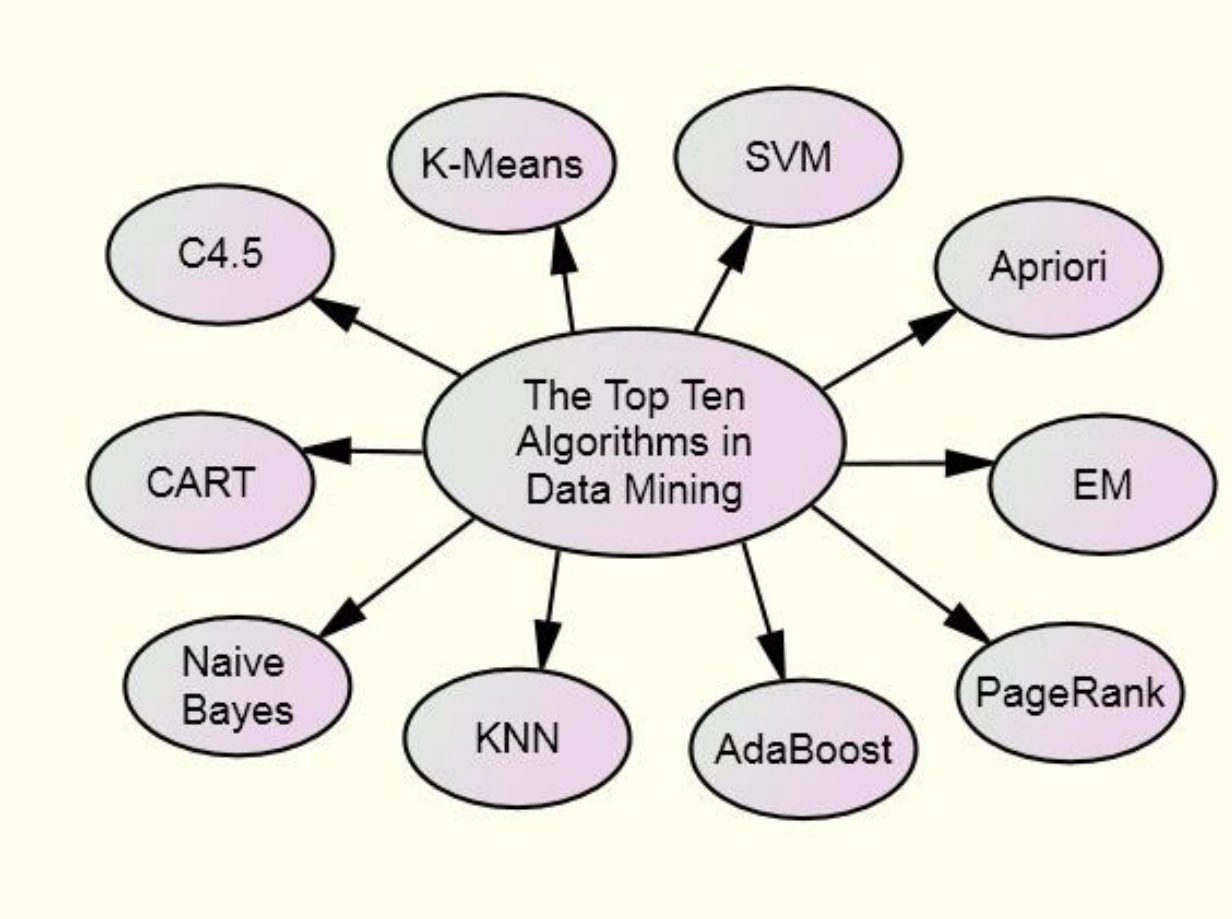
- 分布式文件系统
- 分布式数据库系统
- 分布式邮件系统

谷歌机房

四、数据挖掘

1. 基本概念

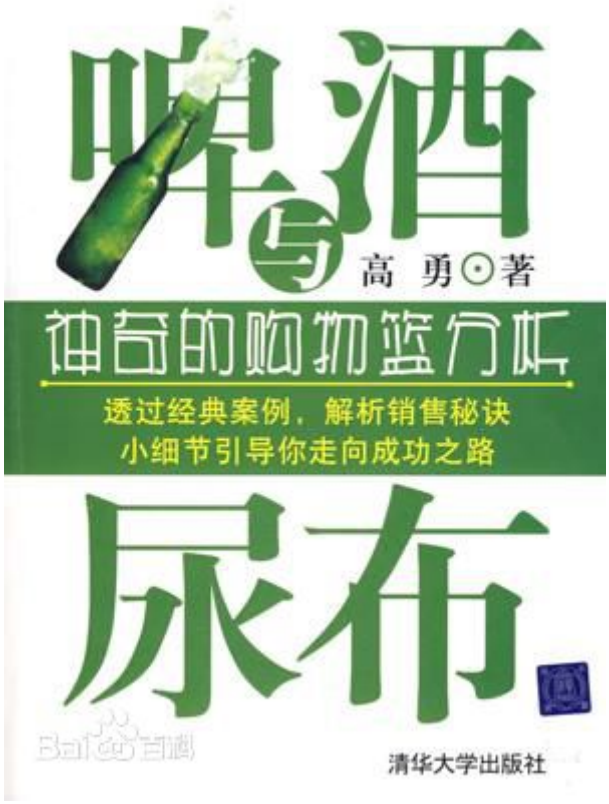
> 数据挖掘一般是指从大量的数据中通过算法搜索隐藏于其中信息的过程。数据挖掘通常与计算机科学有关，并通过统计、在线分析处理、情报检索、机器学习、专家系统（依靠过去的经验法则）和模式识别等诸多方法来实现上述目标。



数据挖掘十大经典算法

2. 主要应用

- 分类算法
- 聚类算法
- 关联规则
- 协同过滤



关联规则经典案例

五、集群部署

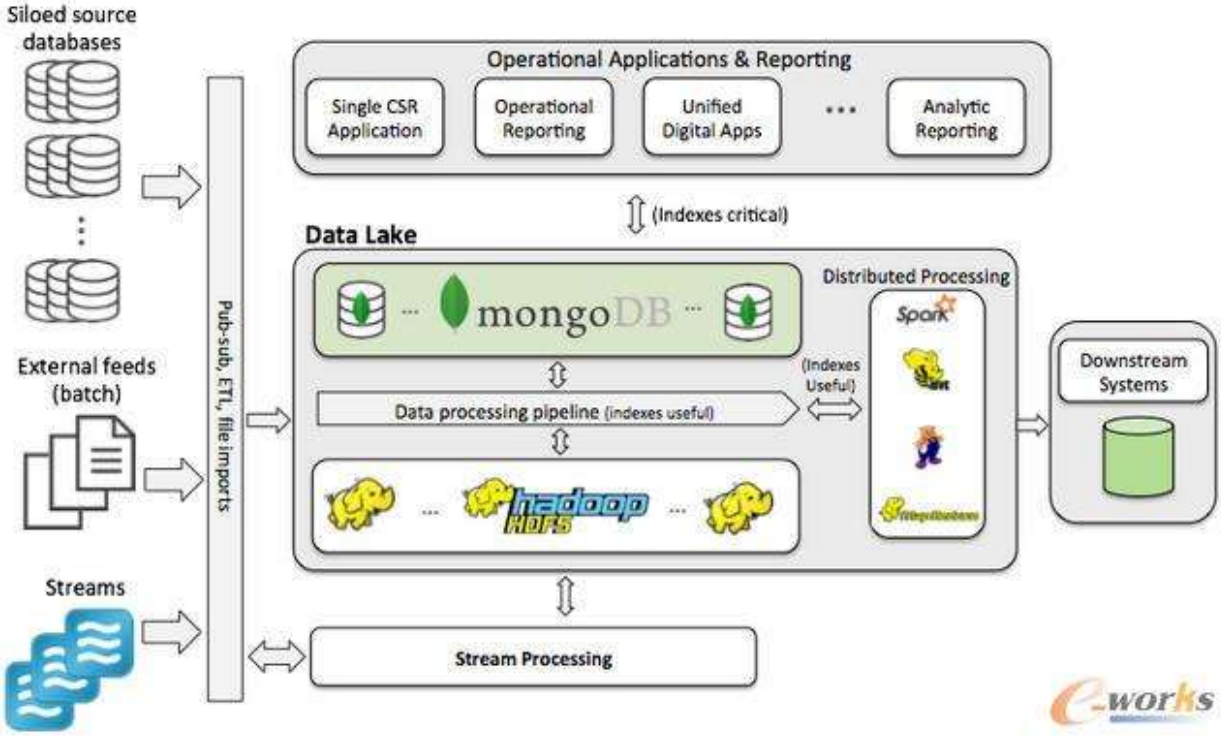
> 服务器集群就是指将很多服务器集中起来一起进行同一种服务，在客户端看来就像是只有一个服务器。集群可以利用多个计算机进行并行计算从而获得很高的计算速度，也可以用多个计算机做备份，从而使得任何一个机器出现问题整个系统还是能正常运行。

阿里巴巴数据中心

六、大数据处理

1. 大数据体系

处理大数据的软件及组件，构建一个完整的大数据生态圈。



大数据生态圈

2. 主流软件

- Hadoop

> Hadoop的框架最核心的设计就是：HDFS和MapReduce。HDFS为海量的数据提供了存储，则MapReduce为海量的数据提供了计算。

- Hive

> 基于Hadoop的一个数据仓库工具，可以将结构化的数据文件映射为一张数据库表，并提供简单的sql查询功能，可以将sql语句转换为MapReduce任务进行运行。 可以通过类SQL语句快速实现简单的MapReduce统计，不必开发专门的MapReduce应用，十分适合数据仓库的统计分析。

- Spark

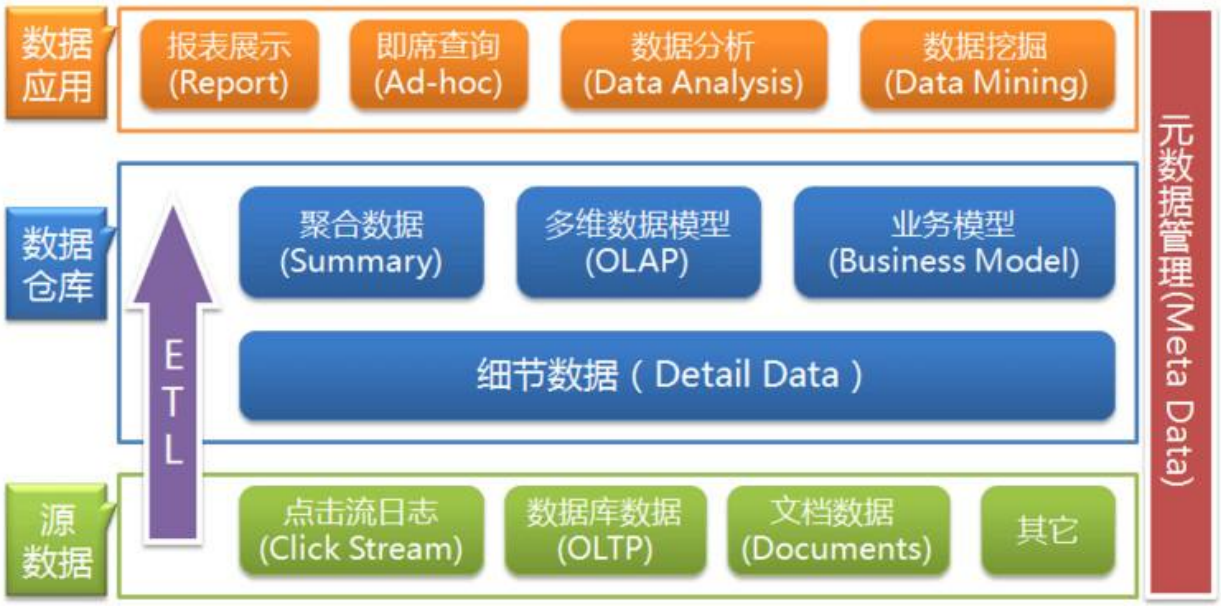
> 创建 Spark 是为了支持分布式数据集上的迭代作业，但是实际上它是对 Hadoop 的补充，可以在 Hadoop 文件系统中并行运行。

- Sqoop

> 一款开源的工具，主要用于在Hadoop(Hive)与传统的数据库(mysql、postgresql…)间进行数据的传递，可以将关系型数据库（例如 ： MySQL ,Oracle ,Postgres等）中的数据导入到HDFS中，也可以将HDFS中的数据导入到关系型数据库中。

七、数据仓库

> 数据仓库中的数据是在对原有分散的数据库数据抽取、清理的基础上经过系统加工、汇总和整理得到的，必须消除源数据中的不一致性。



数据分析流程图

> 数据仓库的数据主要供企业决策分析之用，所涉及的数据操作主要是数据查询，一旦某个数据进入数据仓库以后，一般情况下将被长期保留，也就是数据仓库中一般有大量的查询操作，但修改和删除操作很少，通常只需要定期的加载、刷新。

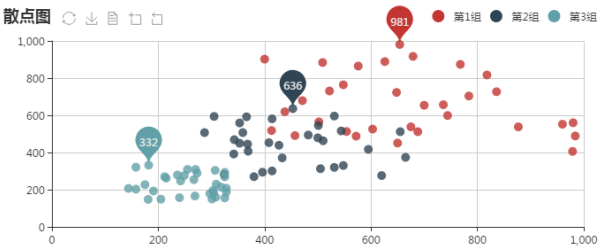
八、数据可视化

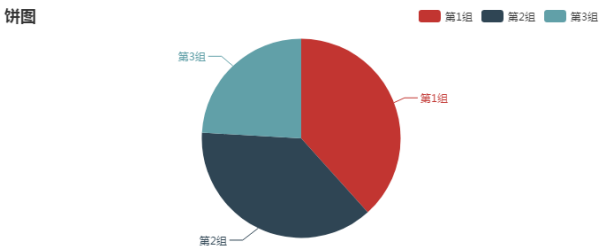
> 利用图形、图像处理、计算机视觉以及用户界面，通过表达、建模以及对立体、表面、属性以及动画的显示，对数据加以可视化解释。

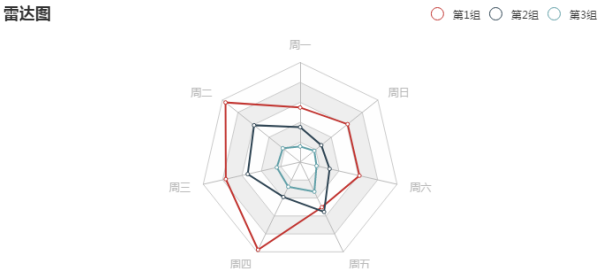
1. Echarts（Enterprise Charts商业级数据图表）

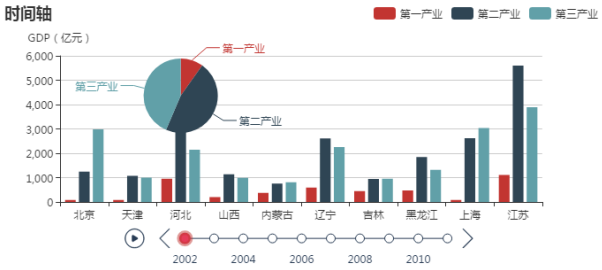
> 百度开源项目，最初为了摆脱Flash，满足各商业体系的报表需求，现在已可应用于BI及数据可视化领域。

2. Echarts-效果图





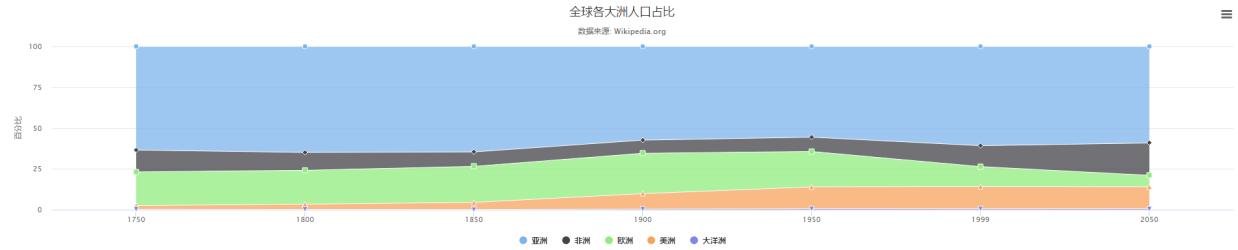


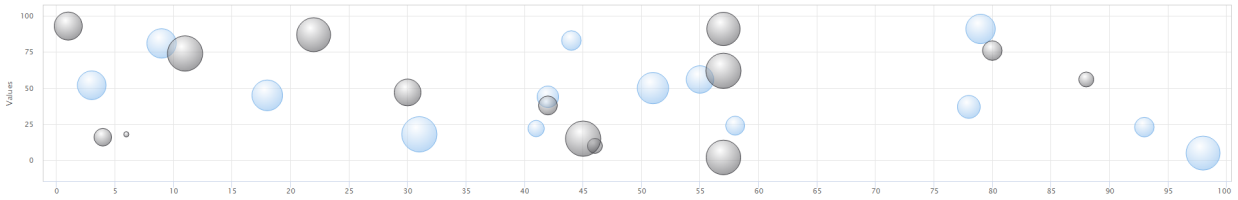


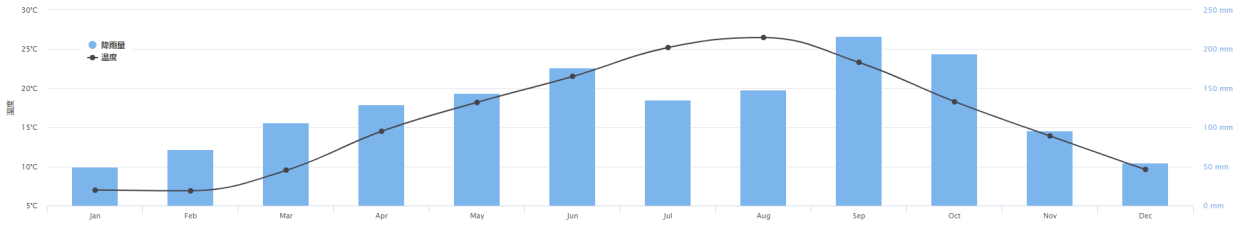
3. HighCharts

> Highcharts 是一个用纯JavaScript编写的一个图表库， 能够很简单便捷的在web应用程序添加有交互性的图表。

4. HighCharts-效果图


全球各大洲人口占比
数据来源: Wikipedia.org
亚洲
非洲
欧洲
美洲
大洋洲




降雨量
温度

5. 项目效果图


6个监控区域
0个区域告警
监控区域
市一区域
妇保区域
音乐喷泉区域
热力分布
数据分析
示意图
热力图
迁徙图
异常事件
四川 0.7 %
山东: 6.2 %
安徽: 8.5 %
福建: 12.8 %
江苏: 45.0 %
江西: 26.8 %