极光数据管理系统建议书

课 程：信息存储与管理

年 级：2017级软件工程

小组成员：王 晋 10161511427

罗亦翔 10175101477

张 丽 10175101176

**目录**

[**极光数据管理系统建议书 1**](#_Toc21353292)

[**一、背景及意义 3**](#_Toc21353293)

[**二、数据管理问题 3**](#_Toc21353294)

[**三、采用的数据库 4**](#_Toc21353295)

[**四、数据源 4**](#_Toc21353296)

[**五、技术路线 6**](#_Toc21353297)

## 背景及意义

极光是自然界的一种绚丽多彩的发光现象，其发生是由于太阳带电粒子流进入地球磁场，在地球南北两极附近地区的高空，夜间出现的灿烂美丽的光辉。极光是一种等离子体现象，主要发生在具有磁场的行星上的高纬度区域，而在地球上的极光带即是经度上距离地磁极10°至20°，纬度宽约3°至6°的区域。早在我国先秦时代的著名古籍《山海经》中，便已经存在关于极光的记载。书中写到一位钟山之神，名字叫做烛龙，它“人面蛇身，赤色，身长千里，钟山之神也。”相貌有如一条红蛇，能在夜空中闪闪发光。广大研究者认为，文中所说烛龙，实际上就是极光。

极光作为唯一可以直接用肉眼观察到的地球物理现象，是研究日地关系的一个重要课题。也正因为如此，它与地球物理学、天体物理学以及大气物理学有着紧密联系。记录、分析、研究各个时期极光的变化与特点，有助于我们了解不同时期的太阳风暴等太阳活动以及其对地球的影响方式与程度、地球磁场以及磁层-电离层的相互作用、高空电离层以及天气等等的变化规律，尤其是和其他学科，如考古地磁学，一起进行历史上地球磁极位置变迁的探讨，因此有着重要的研究意义，更加应当丰富对极光数据的收集与管理。这些数据对于现代科学研究有一定的参考价值。事实上，它已成为世界上宝贵的科学财富。

## 数据管理问题

首先，极光数据是一种海量、巨量资料，这些资料多来源于地球南北两极附近地区，我国的主要依靠位于北极的黄河站观测并收集海量的极光数据，仅仅北极黄河站一年之内收集的图像数量约接近一百万张。这些数据具有大数据基本的 4V 特性（数量Volume，种类Variety，速度Velocity，价值Value）。由于极光分类数据具有多源的特性，极光形态学研究的 4V 特性可以引发新型面向极光分类研究的海量数据管理和分析技术。在海量数据中准确、快速、自动地检测极光事件是进一步研究极光物理机制的基础，是有效利用日积月累产生的海量极光数据面临的首要问题。为了检测极光活动的变化，快速浏览极光活动的变化情况，空间物理研究人员通常采用的一个方法是制作Keogram图。它是从一段的全天空极光视频的每一帧极光图像中抽取沿地磁南北方向一条极光亮度信息，以时间为序列排列成极光活动图。显然，Keogram图方法丢失了极光二维的形态和运动的大量信息，不能完整地反映复杂多变的极光活动。通过Keogram图浏览极光活动，会错过许多重要的极光形态和运动信息，例如极光弧上小尺度涡旋结构，以及发生在全天空图像天顶南北磁子午线以外的极光事件等等。[1]

在本极光数据管理项目中，极光数据主要包括以下五类：1. 极光数据元信息；2. 极光数据图像信息；3. 极光数据缩略图信息；4. 用于生产Keogram的条列信息；5. 从极光图像提取出来用于相似性比较的特征信息。极光数据的特征有以下几点：极光数据主要以img格式的形式存在，需要对极光数据进行预处理后，提取出极光数据元信息，并生成极光数据图像信息，极光数据缩略图信息，用于生成Keogram的条列信息以及用于相似性比较的特征信息。

基于极光数据的以上特性，面对日积月累产生的海量极光数据，尤其是伴随着信息技术和数字相机技术的快速发展，极地科学研究见证了极光数据的指数增长，在极光数据存储与管理时需要选择更有利于数据的传递和交互的数据格式与数据库。我们应当考虑的首要问题是尽可能快速地发现极光现象的发生及活动特征，从而进一步研究极光的物理机制及相关动力学过程的。因此，极光数据管理项目一大重要问题是如何为研究极光现象提供有效的自动化分析手段，从而能够提供充足而有效的事件用于统计学分析。

## 三、采用的数据库

非关系型数据库主要有四种数据存储类型：键值对存储（key-value），文档存储（document store），基于列的数据库（column-oriented），还有就是图形数据库（graph database）。目前主流的数据模型有**HBASWE、MongoDB、Redis、Neo4J。**

**Redis**是一个开源的使用ANSI C语言编写、支持网络、可基于内存亦可持久化的日志型、Key-Value数据库，并提供多种语言的API。它适用于数据变化较少，执行预定义查询，进行数据统计的应用程序。但是数据库容量受到物理内存的限制,不能用作海量数据的高性能读写,因此Redis适合的场景主要局限在较小数据量的高性能操作和运算上，不适合做存储和分析，所以不适合极光系统需要大量存储的需求。

**HBASE**是一个建立在HDFS之上的分布式，提供高可靠性，高性能，列存储，可伸缩，实时读写NoSQL的数据库系统。但它不支持条件查询，只有简单的字符类型。不支持二级索引，只支持一个大的primary id。而极光系统需要多级索引做过滤条件，从而查询相应的数据，如需要根据给定的时间段、波段、极光类型，查找相应的极光图片并展示，并根据给定的图片，给出现有图片中与上传图片最相似的图片，展示图片并给出极光名称以及相关数据等，上述查询显然HBASE的不支持二级索引，不支持条件查询已不太合适，而MongoDB支持的查询语言非常强大，可以支持复杂的数据结构，不受二维表形式的限制，可以实现类似关系数据单表查询的绝大部分功能。

**Neo4J**是一个高性能的,NOSQL图形数据库，它将结构化数据存储在网络上而不是表中，Neo4j也可以被看作是一个高性能的图引擎，该引擎具有成熟数据库的所有特性。而它不适合记录大量基于事件的数据，极光系统的数据包括：极光数据元信息，极光数据图像信息，极光数据缩略图信息，用于生产Keogram的条列信息，以及从极光图像提取出来用于相似性比较的特征信息。除了海量的原始极光数据，还有海量的分类实验数据，比如人工标注数据等等。这些数据需要大量的内存，但Neo4J适合图形一类的数据，如社会关系，公共交通网络，地图及网络拓谱等，所以不太适合。

**MongoDB**是面向文档、模式自由的文档型数据库。它经常被誉为具有RDBMS功能的NoSQL数据库，还带有交互式shell。

1. 更高的写负载，MongoDB拥有更高的插入速度
2. 文档结构的存储方式，能够更便捷的获取数据。MongoDB支持二维空间索引，比如管道，因此可以快速及精确的从指定位置获取数据。MongoDB在启动后会将数据库中的数据以文件映射的方式加载到内存中。如果内存资源相当丰富的话，这将极大地提高数据库的查询速度。
3. 性能优越。
4. 高效的传统存储方式：支持二进制数据及大型对象（如照片或图片），我们的极光系统就需要存储大量的图片，所以非常适用。
5. 数据可扩展性：MongoDB提供基于Range的Auto Sharding机制：一个collection可按照记录的范围，分成若干段，切分到不同的Shard上。Shard机制实现了数据规模的可扩展性，随着数据规模增加，可以通过增加Shard来满足需求。

综上所述，总结出MongoDB更适合极光数据管理系统的原因：

1. 极光数据的特点是数据量大，需要存储大量的图片，而MongoDB的优点之一就是高效的存储方式，支持二进制数据及大型对象（如照片或图片）
2. MongoDB支持复杂的查询，可以通过字段，范围查询进行搜索，并且还支持正则表达式搜索，不受二维表形式的限制。这满足了极光系统需要根据给定的时间段、波段、极光类型，查找相应的极光图片和根据给定的图片查找相似的图片等功能。
3. 极光系统采用数据访问对象模式，用于把低级的数据访问API或操作从高级的业务服务中分离出来，MongoDBCompassCommunity为MongoDB可视化管理工具，可用它来对极光数据进行管理。

## 四、数据源

极光数据主要包括：极光数据元信息，极光数据图像信息，极光数据缩略图信息，用于生产Keogram的条列信息，以及从极光图像提取出来用于相似性比较的特征信息。极光数据主要以img格式的形式存在，需要对极光数据进行预处理后，提取出极光数据元信息，并生成极光数据图像信息，极光数据缩略图信息，用于生成Keogram的条列信息以及用于相似性比较的特征信息。

## 五、技术路线

1. **数据介绍**

空间环境观测系统。该观测系统运用了三台完全一样的单色全天空数字 CCD 相机。这些仪器在极夜的环境下，可以 24 小时以 10 秒为间隔连续产生 3 个波段(427.8, 557.7 和 630.0 nm)的极光数据。

从此可以看出极光研究是海量数据的研究。除了海量的原始极光数据，还有海量的分类实验数据，比如人工标注数据，特征提取算法和参数，分类算法和参数，实验结果，模型数据以及极光探索分析产生的可视化数据。

1. **功能设计**

**案例系统中包括：**

数据导入——上传相应的极光数据文件，进行预处理后将数据存入数据库中；

数据搜索——根据给定的时间段、波段、极光类型，查找相应的极光图片并展示；原图查询，根据给定的图片，给出现有图片中与上传图片最相似的图片，展示图片并给出极光名称以及相关数据；

人工标注——查找图片后可以通过人工观察，对极光进行相关信息的标注。

**针对系统，我们对功能进行添加以及优化：**

数据导入——可以将极光数据进行分类，上传时选择可创建或删除相应的极光集合，以便区别相似类型的极光。

下载数据——根据用户所选极光下载相关数据和图片。

图像分析——根据用户所选则的极光，展示已经存储极光图像序列，由极光图像中地磁南北方向的线按照时间排列得到时间/纬度图。

1. **技术设计**

根据系统的方案设计，对于数据交换，数据处理，数据分析等。项目根据极光数据的相关特性，选择使用非关系数据库MongoDB，并根据MongoDB的数据库特性，案例中对极光数据库进行了设计：

Database: Aurora:{

“Info”:{

“tag”,

“name”,

“time”：{

“year“

“month”

“day”

“hour”

“minute”

“second”

}

“band”,

“location”:{

“obstation”

“Latitude”

“Longitude”

}

“manual type”

“feature”

}

“Image”:{

“files”:{

“img”

“Keogram”

…

}

}

用户视图方面·，项目选择使用JavaScript、CSS以及HTML实现基于网页的用户视图，将数据完整、清晰地展示在系统网页中，并便于用户对数据进行管理和执行相关操作。 后端使用Java语言与非关系数据库MongoDB进行链接。