TECHNOLOGY SUMMIT 5.0
DIGITAL • MOBILE • GLOBAL

Enterprise Search

Sam Beran
sam.beran@pearson.com

Who Am I?

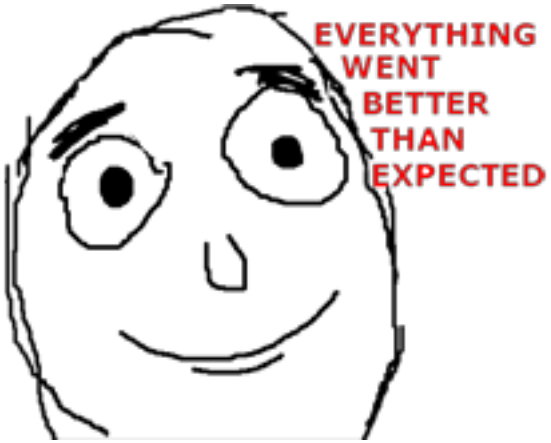
A Story You Have Heard Before

A Story You Have Heard Before

Shiny New Site!

Students

First Name	Last Name	Grade	Test
Jim	Johnson	8	Math
Jane	Li	6	Science
Karen	Anderson	9	Reading
Luke	Skywalker	4	Science



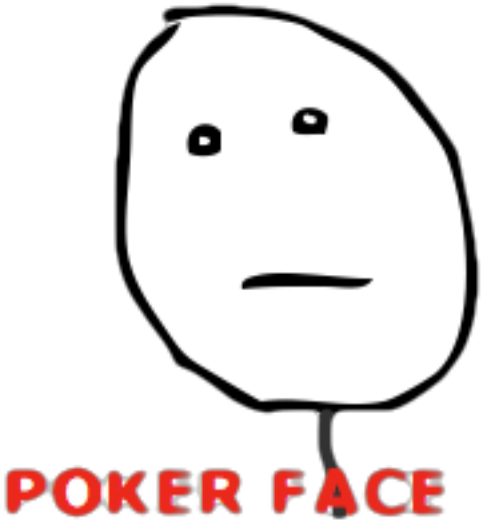
Version 2 - Your Designer Gets Clever

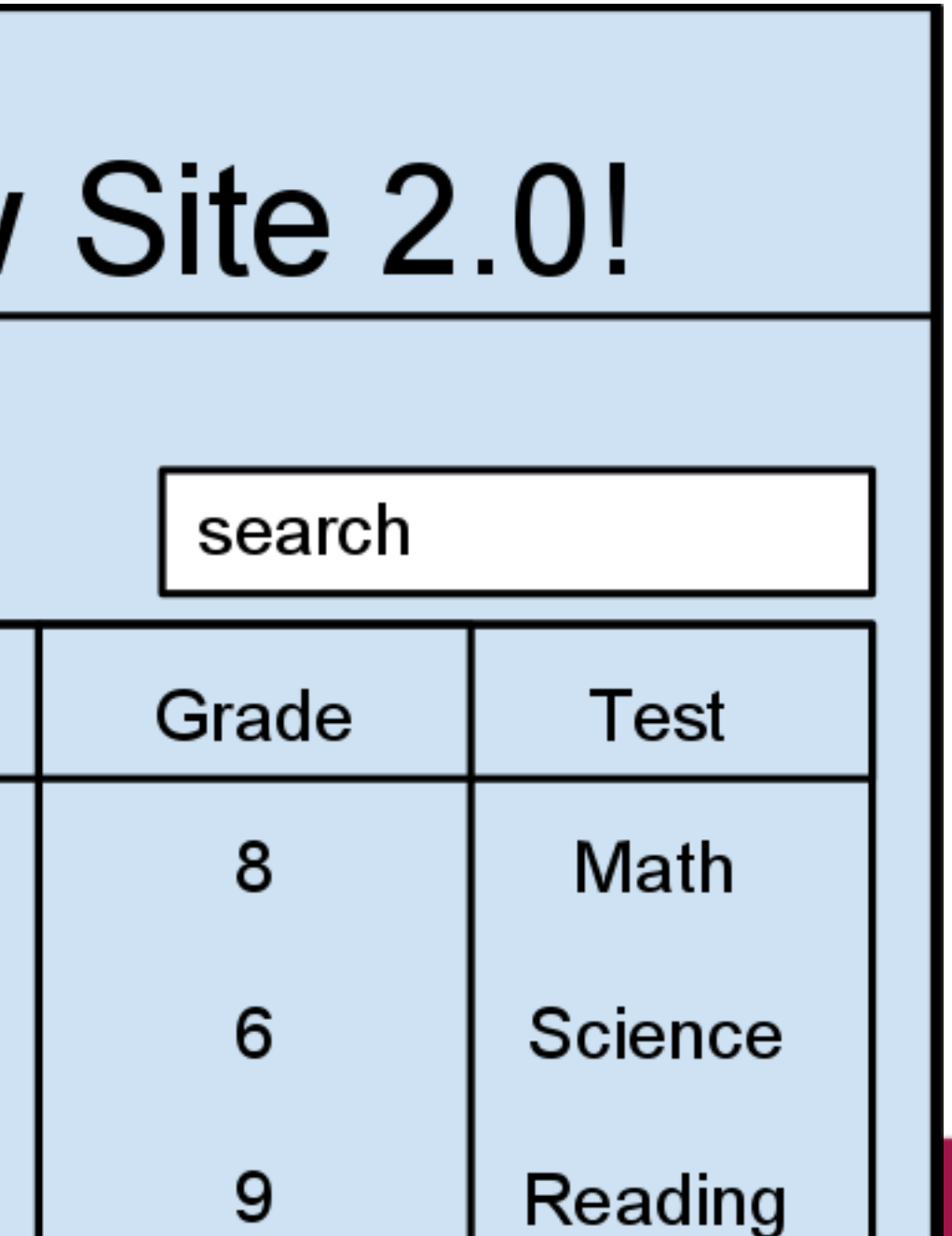
Shiny New Site 2.0!

Students

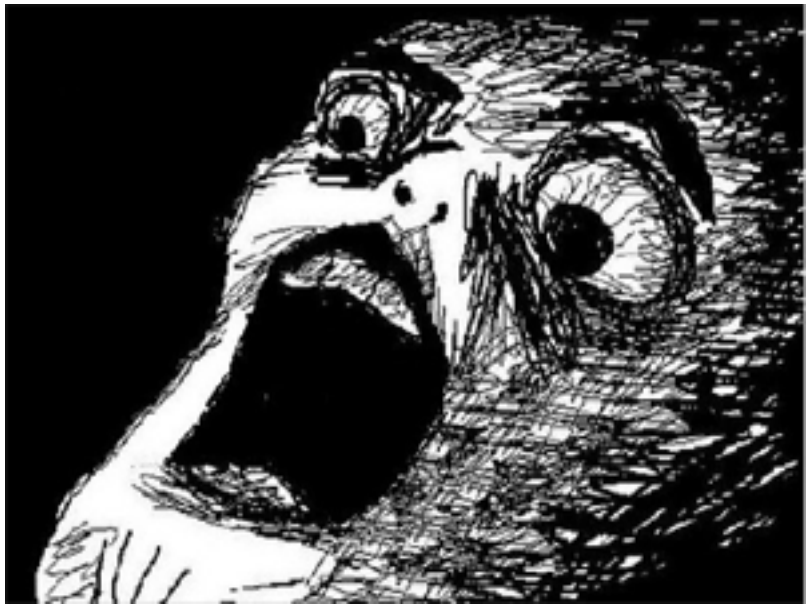
search

First Name	Last Name	Grade	Test
Jim	Johnson	8	Math
Jane	Li	6	Science
Karen	Anderson	9	Reading
Luke	Skywalker	4	Science





Looking at the requirements,



I need SEARCH

"I'll Use A Database Query!"

```
select * from students where
firstname like '%term%' or
lastname like '%term%' or code
like '%term%' order by lastname
```

select * from students where firstname like '%foo%' or lastname like '%foo%' or code like '%foo%' order by lastname

You will have to parse the input
"joe smith" => "joe", "smith"

Full Table Scan

Best results will not appear first

Wrong tool for the job.

DB FULLTEXT index?

Oracle, SQLServer, PostgreSQL, MySQL

Oracle, SQLServer, PostgreSQL, MySQL

FULLTEXT not supported on InnoDB

Database FULLTEXT is a pretty good option.

Database FULLTEXT is a pretty good option.

Avoid sync issues

Database FULLTEXT is a pretty good option.

Avoid sync issues

Little additional complexity

Database FULLTEXT is a pretty good option.

Avoid sync issues

Little additional complexity

Few advanced features

The Cadillac of Search

Based on lucene

Lucene

- Low level Search API
- 100% Java
- Depends on local index files
- Esoteric API

Features of Solr

HTTP api

localhost:8983/solr.war/sele

localhost:8983/solr.war/select/?q=electronics&indent=on


```
▼<result name="response" numFound="3" start="0">
  ▼<doc>
    ▼<arr name="cat">
       <str>electronics</str>
       <str>memory</str>
     </arr>
     <str name="id">VS1GB400C3</str>
     <bool name="inStock">true</bool>
     <str name="manu">Corsair Microsystems Inc.</str>
     <date name="manufacturedate_dt">2006-02-13T15:26:37Z</date>
    ▼<str name="name">
       CORSAIR ValueSelect 1GB 184-Pin DDR SDRAM Unbuffered DDR 400 (PC 3200) System Memory - Retail
     </str>
     <str name="payloads">electronics|4.0 memory|2.0</str>
     <int name="popularity">7</int>
     <float name="price">74.99</float>
     <str name="store">37.7752,-100.0232</str>
   </doc>
  ▼<doc>
    ▼<arr name="cat">
       <str>electronics</str>
       <str>memory</str>
     </arr>
    ▼<arr name="features">
       <str>CAS latency 3, 2.7v</str>
     </arr>
     <str name="id">VDBDB1A16</str>
     <bool name="inStock">true</bool>
     <str name="manu">A-DATA Technology Inc.</str>
     <date name="manufacturedate_dt">2006-02-13T15:26:37Z</date>
    ▼<str name="name">
```

JSON Support

http://localhost:8983/solr.war/select/?q=electronics&indent=on&wt=json

{ "response":
{"numFound":3,"start":0, "docs":
[
{ "id":"VS1GB400C3", "name":"CORSAIR ValueSelect 1GB"
"manu":"Corsair Microsystems Inc.",
"price":74.99, "popularity":7, "inStock":true,
"store":"37.7752,-100.0232", "manufacturedate_dt":"2006-02-13T15
26:37Z", "payloads":"electronics|4.0 memory|2.0", "cat":["electronics","
memory"]},

Spellcheck

http://localhost:8983/solr/spell?q=delll&spellcheck=true&spellcheck.collate=true


```
<lst name="spellcheck">
  <lst name="suggestions">
    <lst name="delll">
      <int name="numFound">1</int>
      <int name="startOffset">0</int>
      <int name="endOffset">5</int>
      <arr name="suggestion">
          <str name="collation">dell ultrasharp</str>
    </lst>
</lst>
```

Suggestions

http://localhost:
8983/solr/suggest?q=ac

<?xml version="1.0" encoding="
UTF-8"?> <response> <lst
name="spellcheck"> <lst name="
suggestions"> <lst name="ac">
<int name="numFound">2</int>
<int name="startOffset">0</int>
<int name="endOffset">2</int>
<arr name="suggestion">
<str>acquire</str>
<str>accommodate</str> </arr>
</lst> <str name="collation"
>acquire</str> </lst> </lst>

Highlighting

http://localhost:8983/solr/select/?q=corsair&fl=name,id&hl=true&hl.fl=name,features


```
<lst name="highlighting">
  <lst name="VS1GB400C3">
    <arr name="name">
      <str>
<em>CORSAIR</em>
ValueSelect 1GB 184-Pin DDR
SDRAM Unbuffered DDR 400
(PC 3200) System Memory -
Retail
      </str>
    </arr>
  </lst>
```

Facet Query

http://localhost:8983/solr.war/select/?q=Corsair&facet=true&facet.field=inStock

```
<lst name="facet_counts">
  <lst name="facet_fields">
    <lst name="inStock">
      <int name="true">2</int>
    </lst>
  </lst>
</lst>
```

Advanced Queries

Search Specific Fields

firstName:John lastName:Smith

Required / Ignored Terms

division +math -biology

Exact Match

"C++ Vector"

Boolean Logic

(firstName:Jennifer AND lastName:Smith)

Sharding / Replication

DEMO

~ 15 minutes

Downsides

SPOT Principle
No Single Point Of Truth

SPOT Principle

Your data is now in two places.

SPOT Principle

Your data is now in two places.

Partial Solution: Return IDs from Solr

SPOT Principle

Your data is now in two places.

Partial solution: return ID from Solr

Look up data from DB

Frequent commits == bad performance

Frequent commits == bad performance

use commitAfter

Frequent commits == bad performance

use commitAfter

queue commits

Solr 4.0 - NRT

Solr 4.0 - NRT

Soft commit - real time updates

You know, for search.

- New kid on the block
- Real time updates
- No XML
- No schema

Questions?

Further Reading

- http://lucene.apache.org/solr/
- http://wiki.apache.org/solr/
- http://www.elasticsearch.org/
- https://github.com/tjake/Solandra