# data, OLAP
# and some architecture thoughts.

memorphic

date :

20 june 2008

a presentation to :

LFPUG

# purpose of this presentation

Different architecture models for online/offline applications

Overview of OLAP

Understanding OLAP's limitations and seeing where SQLite can assist

# the problems we face today with most applications

Either lots of querying which can get frustrating or one long download when you start the app

Application lockup either in parsing, rendering or calculation easily solved using asynchronous callbacks like setTimeout, Timer or Enterframe

No offline capability

Common data formats we use today
XML
AMF
JSON
(REST with custom formats)
With AIR
+=SQLITE

# Architecting for online offline systems

## and ideally everything in between at the same time

online

caching strategies or repeated querying

memory consumption

speed of queries

Any user driven drill down is thus slow

offline

requires install and thus hassle to deploy in large corperates

memory consumption might still be a problem

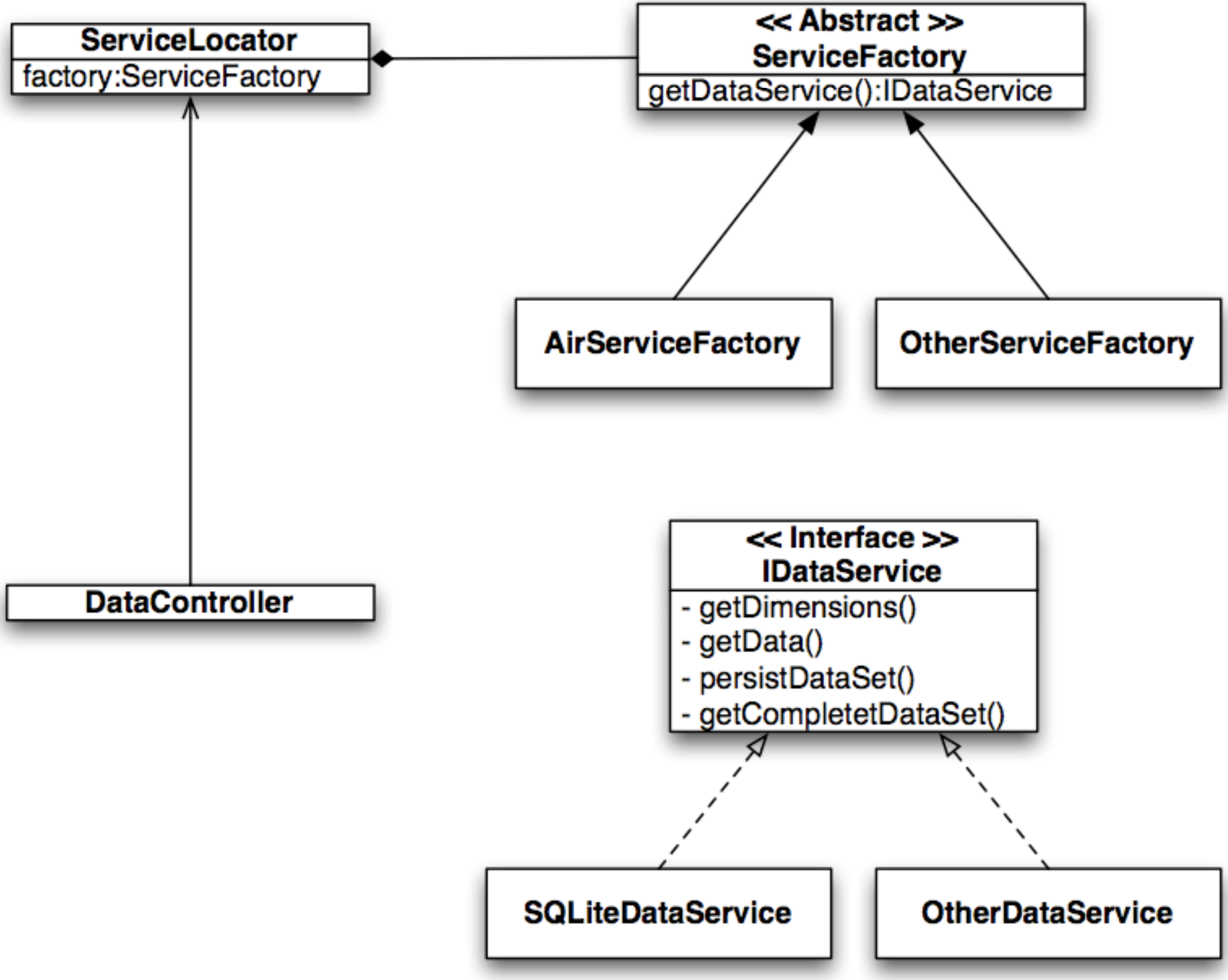
developing for both

the only difference need be the entry-point
(mx:Application vs mx:WindowedApplication)

forces you to develop around interfaces (this is a good thing!)

service locator on which the factory gets set by the entry point

show some code.


ServiceLocator
factory:ServiceFactory
<< Abstract >>
ServiceFactory
getDataService():IDataService
AirServiceFactory
OtherServiceFactory
DataController
<< Interface >>
IDataService
- getDimensions()
- getData()
- persistDataSet()
- getCompletetDataSet()
SQLiteDataService
OtherDataService

# back to some data stuff

knowing when to use what format and carrier is 99% of the secret

AMF

- ☐☐whenever you deal with a fair amount of data over the net - online or off

XML

- never! small datasets, config files when working with SOAP a wrapper is advised which speaks AMF

SQLite

- large reports you have to drill into
- local caching
- always running

Some tips:

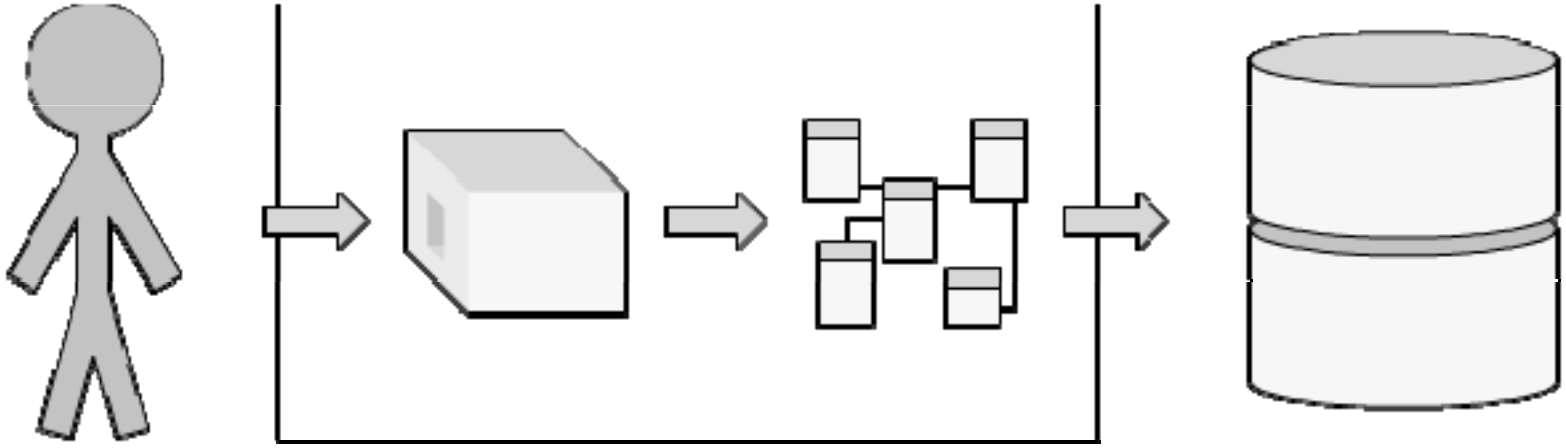
- use transactions for lots of local operations
- build database on server and download

Sockets or FLV metadata

- tickers, large graphs that feed and missing a beat is not critical

# what is OLAP?

Short for Online Analytical Processing, OLAP is a set of tools used for analysing different dimensions in a database.



# OLAP terms

Cube : OLAP cubes can be thought of as extensions to the two-dimensional array of a spreadsheet

Dimension **:** A collection of data of the same type, allowing us to structure the multidimensional database. A dimension is sometimes referred to as an axis. In a measure, each cell of data is associated with one single position in each dimension. Time, Location and Product are the classic dimensions.

Hierarchy **:** The positions of a dimension organized according to a series of cascading one to many relationships. This way of organizing data is comparable to a logical tree, where each member has only one parent but a variable number of children.

Attribute : Also called a dimension member.

Measure : Most often an integer or decimal type, structured by the dimensions. Salary, Value and Quantity are classic measures.

# OLAP operations

Slice : A slice is a subset of a multi-dimensional array corresponding to a single value for one or more members of the dimensions not in the subset.

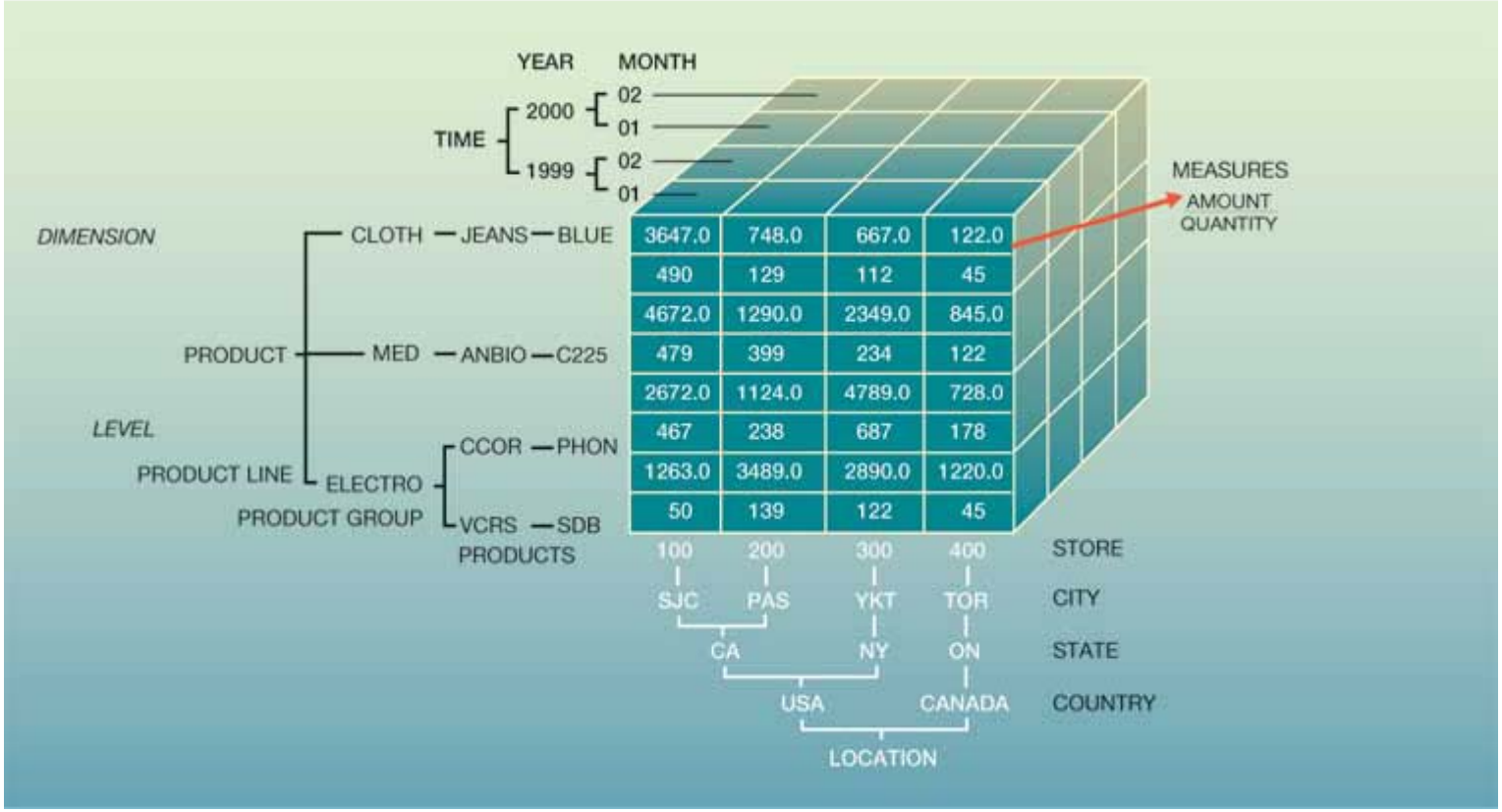
Dice **:** The dice operation is a slice on more than two dimensions of a data cube (or more than two consecutive slices).

Drill Down / Up **:** Drilling down or up is a specific analytical technique whereby the user navigates among levels of data ranging from the most summarized (up) to the most detailed (down).

Pivot : To change the dimensional orientation of a report or page display.

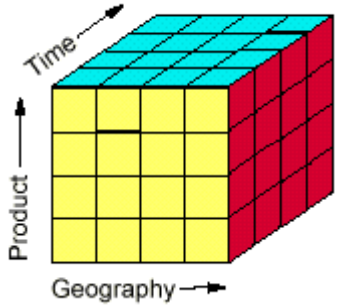
# OLAP operations

Figure 2 OLAP sales cube




http://www.research.ibm.com/journal/sj/414/colos2.jpg

# flat data




| Age | Gender Status | Marital | Country | Zip Code Code | Area | Through Ads? | Purchased ? URL | Time on |
|---|---|---|---|---|---|---|---|---|
| 53 | Male | Married | US | 07074 | 973 | Yes | Yes | 252 |
| 53 | Male | Married | US | 07074 | 973 | Yes | Yes | 252 |
| 41 | Female | Married | US | 07074 | 973 | Yes | Yes | 325 |
| 29 | Male | Single | US | 07074 | 973 | No | No | 460 |
| 21 | Male | Single | US | 07074 | 973 | No | No | 635 |

Then the dimensions (their attributes) and measures are the following

Dimensions

| | | | | |
|---|---|---|---|---|
| Customer | - | Age | Gender | Marital Status |
| Geography | - | Country | Zip Code | Area Code |
| Purchase | - | Through Ads? | Purchased? | |

Measure - Time spent on URL

Thus asking what how long people spent on the URL based on their gender and marital break down would give you the following data set

| Gender | Married | Single | All |
|---|---|---|---|
| Male | 3211 | 9548.5 | 12759.5 |
| Female | 3274.5 | 2107 | 5831.5 |
| All | 6935.5 | 11655.5 | 18591 |

# Our example data Set (1/2)

| Id | Year | Quarter | Week | Sku | Category | Segment | Location | Area |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 4 | 51 | Product-6 | Category-61 | Segment-6 | Location-5-2 | Area-5-2 |
| 2 | 1 | 1 | 3 | Product-1 | Category-12 | Segment-1 | Location-7-0 | Area-7-0 |
| 3 | 2 | 6 | 68 | Product-4 | Category-45 | Segment-4 | Location-2-1 | Area-2-1 |
| 4 | 1 | 4 | 47 | Product-6 | Category-60 | Segment-6 | Location-23- | Area-23-2 |
| 5 | 1 | 2 | 17 | Product-2 | Category-28 | Segment-2 | Location-7-2 | Area-7-2 |
| 6 | 1 | 1 | 12 | Product-2 | Category-28 | Segment-2 | Location-11- | Area-11-2 |

| Town | Volume | Gross | Change | Age_Band | Gender |
|---|---|---|---|---|---|
| Town-5 | 4996 | 57204.2 | 0 | Age 0 - 25 | Female |
| Town-7 | 4995 | 39010.95 | -0.000200200200200 | Age 0 - 25 | Female |
| Town-2 | 4986 | 110240.46 | -0.001805054151624 | Age 50 - 75 | Female |
| Town-23 | 4998 | 159136.32 | 0.002400960384153 | Age 25 - 50 | Female |
| Town-7 | 5000 | 232700 | 0.0004 | Age 0 - 25 | Female |
| Town-11 | 4999 | 464807.02 | -0.000200040008001 | Age 50 - 75 | Male |

# Our example data Set (2/2)

In our example data set
we have 4 dimensions
Time
    Year, Quarter, Week
Product
    Category, Segment, Sku
Location
    Town, Area, Location
Demographics
    Age Band, Gender

Gross,Volume and Change as our measures

These measures follow different aggregation rules
Gross and Volume use SUM while Change uses AVG

# limitations currently in OLAP

Max of 50k records before it becomes REALLY SLOW
this is only with a simple model however, as little as 300 records can bring it down with a complex enough model

No direct SQLite libraries
in an ideal world you would supply a sqlite db as a data source and not have to query it yourself.

Somework arounds involving SQLite
always give the OLAP cube the simplest possible model

always give it the smallest data set

aggregate where possible with SQL

And now for some random talk and explanation regarding OLAP

# some other mutterings

BigTable as a natural OLAP source

The db file as a cube

Local caching built into frameworks

presented by :


Johannes Nel

johannes.nel@memorphic.com

Memorphic SA

20 Drosdty Centre
Bird Street
Stellenbosch, 7600
South Africa

Telephone: +27 (0) 21 8866066

Peter Hall

peter.hall@memorphic.com

Memorphic UK

75 Highlands Heath
Portsmouth Road
London, SW15 3TX
UK

Telephone: +44 (0) 779 1096899