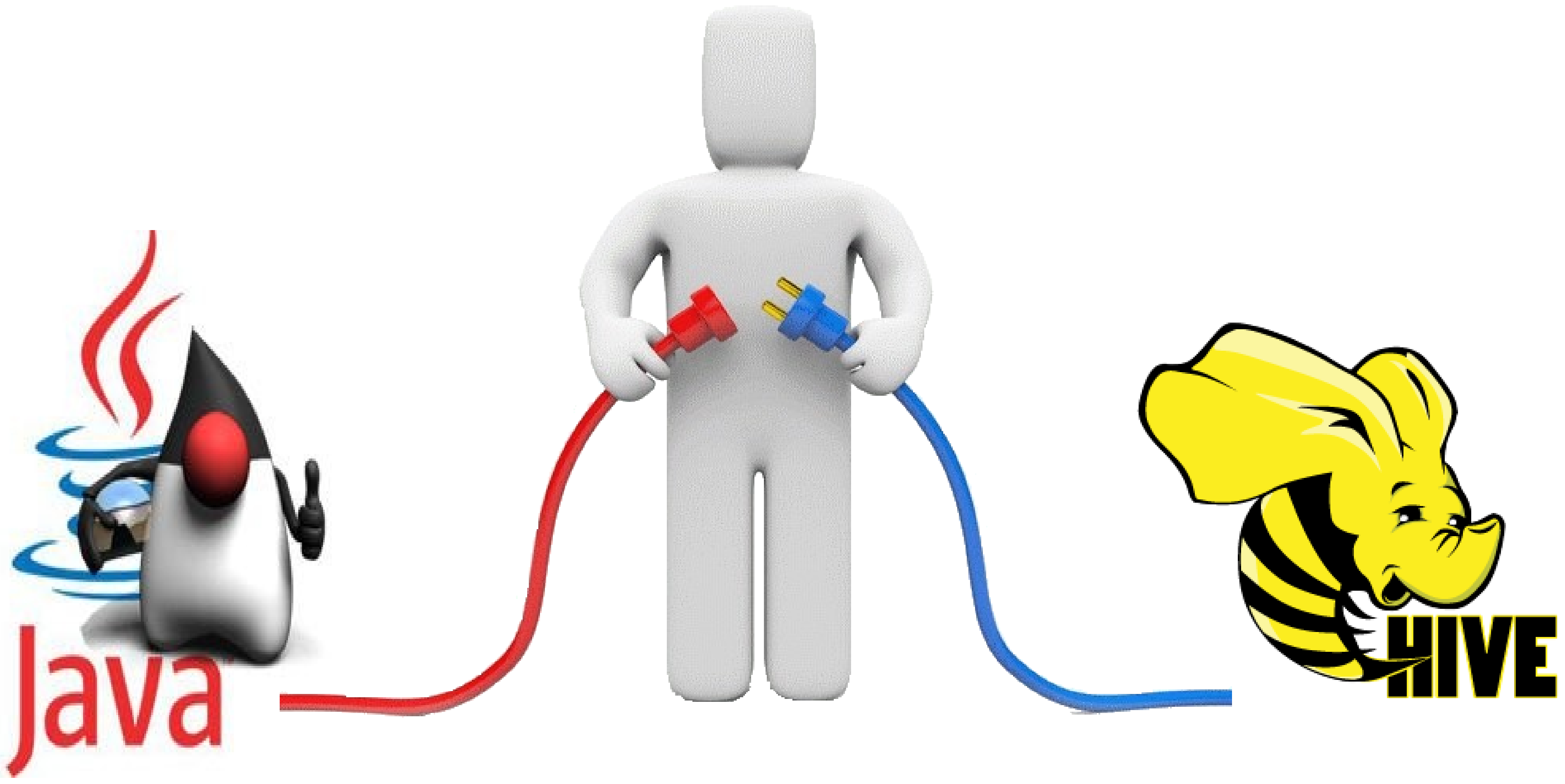
Using jdbc with hive



Agenda

- Who are we
- Jdbc for ETL
- Jdbc for BI people
- Questions

eBuddy

- Web based chat
  - Started in 2003 (no statistics)
  - 1.7B record (im logins)
    - Started basic logging in 2004
- XMS
  - 490M record (xms)
    - Launched May 23, 2011
- Interesting to know
  - Hosting in the US but BI people in Amsterdam
  - Developers are Java centric

ETL

- Connection pooling
  - InitSql (new connections with udfs)

- Jdbc templates (Spring)

ETL – Connection pooling

```
    <bean id="dwhHiveDataSource" class="org.apache.commons.dbcp.BasicDataSource" destroy-
method="close">
        <property name="driverClassName" value="${db.dwh.class}"/>
        <property name="url" value="${db.dwh.url}"/>
        <property name="username" value="${db.dwh.user}"/>
        <property name="password" value="${db.dwh.pwd}"/>
        <property name="minIdle" value="${db.dwh.min}"/>
        <property name="testOnBorrow" value="${db.dwh.testonborrow}"/>
        <property name="validationQuery" value="${db.dwh.validationquery}"/>
        <property name="connectionInitSqls" ref="listInitSql"/>
    </bean>

    <util:list id="listInitSqlLocal">
        <value>SET pool.name=${pool.name}</value>
        <value>SET hive.exec.mode.local.auto=${hive.exec.mode.local.auto}</value>
        <value>add jar ${udf.location}</value>
        <value>create temporary function iptolong as 'com.ebuddy.dwhhive.udf.IpToLong'</value>
    </util:list>
```

ETL - Connection pooling

```
public class ImEtlServiceImpl extends JdbcDaoSupport
```

Then wire it in:

```
 <bean id="im"
class="com.ebuddy.dwhhive.etl.im.ImEtlServiceImpl">
    <property name="dataSource" ref="dwhHiveDataSource"/>
 </bean>
```

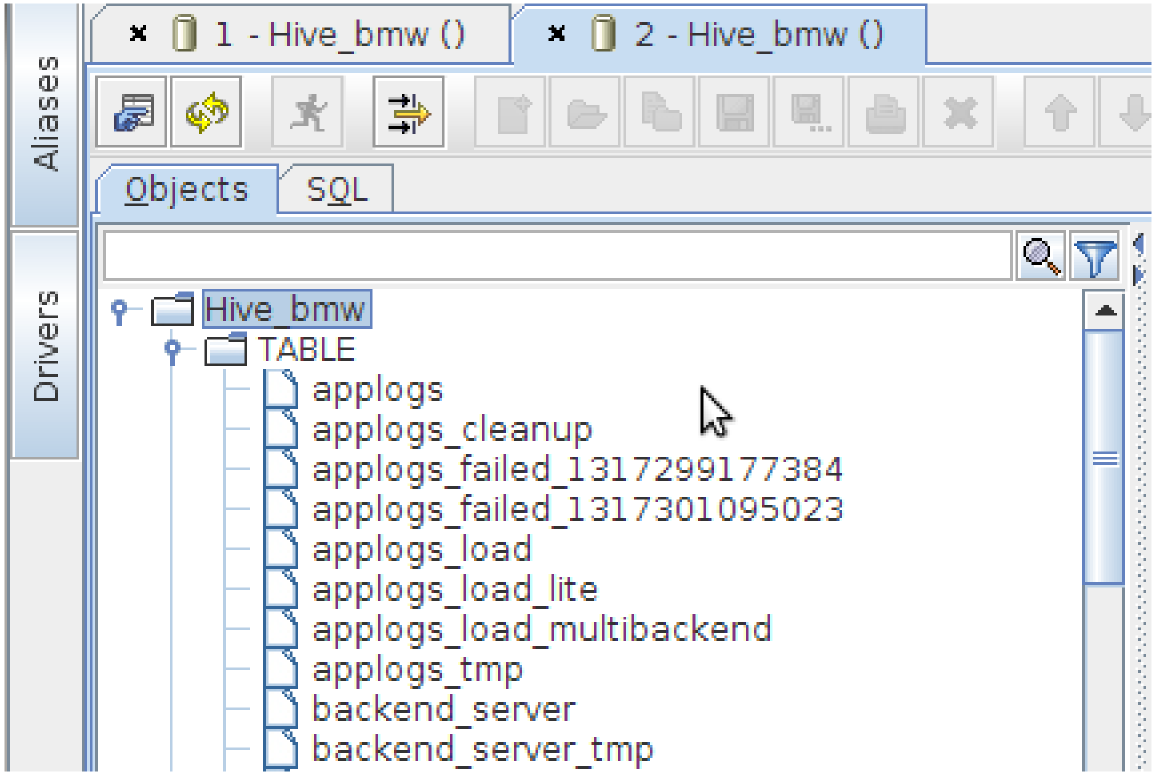
ETL - Jdbc templates

```
String sql = "drop table sometable”;
getJdbcTemplate().execute(sql);

String sql = "select count(*) from sometable”;
long recordCnt = getJdbcTemplate().queryForLong(sql);

String sql = "select * from sometable";
List<Map<String, Object>> recordsLast =
getJdbcTemplate().queryForList(sql);
```

Squirell

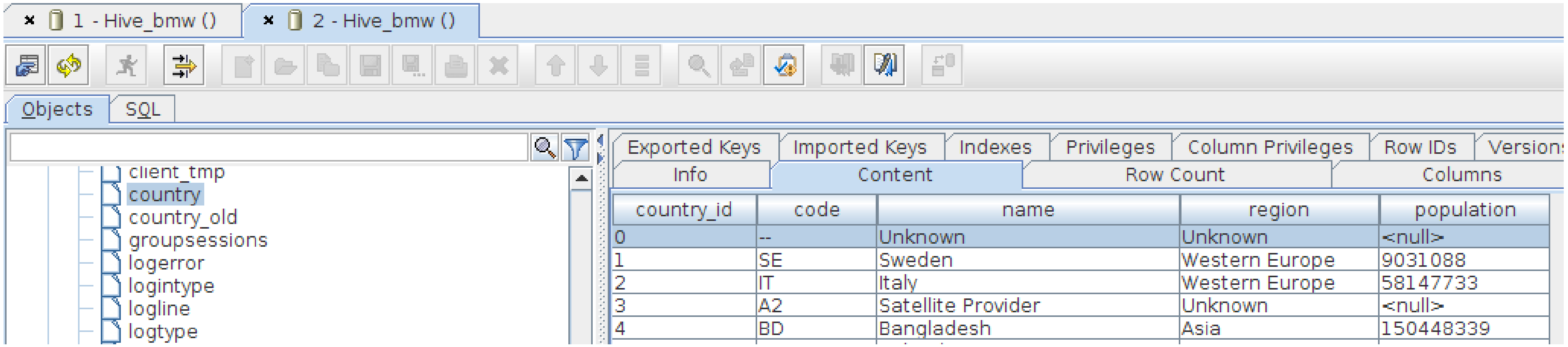
- Meta data (HIVE-1126)
  - List of existing tables/data types
  - Code completion (Ctrl-Space)
- Concurrency
- Performance
  - “Building Output” (HIVE-1815)

Squirell - getTables



Hive_bmw
TABLE
applogs
applogs_cleanup
applogs_failed_1317299177384
applogs_failed_1317301095023
applogs_load
applogs_load_lite
applogs_load_multibackend
applogs_tmp
backend_server
backend_server_tmp

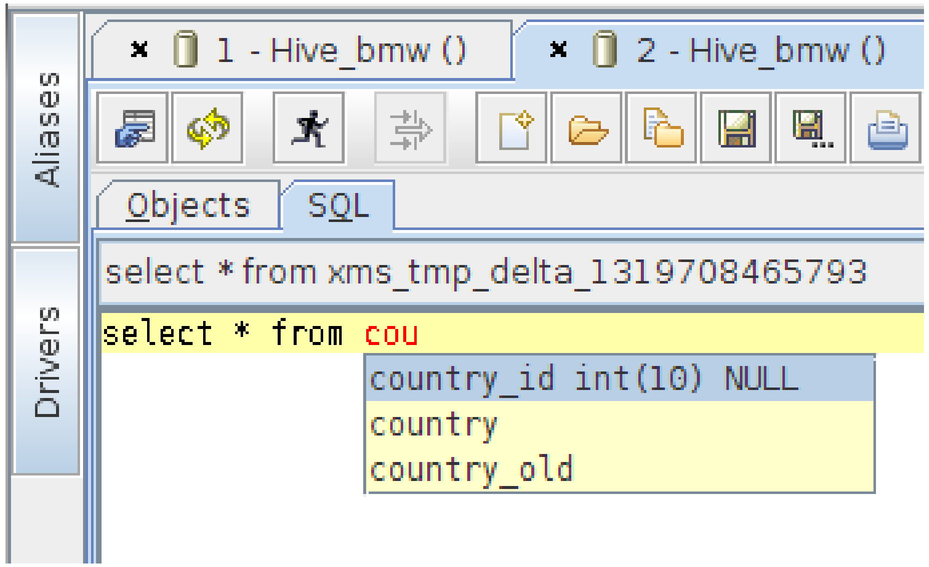
Squirell – View data



client_tmp
country
country_old
groupsessions
logerror
logintype
logline
logtype

country_id	code	name	region	population
0	--	Unknown	Unknown	<null>
1	SE	Sweden	Western Europe	9031088
2	IT	Italy	Western Europe	58147733
3	A2	Satellite Provider	Unknown	<null>
4	BD	Bangladesh	Asia	150448339

Squirell – code completion



```
select * from xms_tmp_delta_1319708465793
select * from cou
country_id int(10) NULL
country
country_old
```

Squirell - “Building Output”

Records	Fetch size	Time on output
1000	1	2min 43sec
1000	50	4sec
1000	100	2sec
1000	1000	1sec

- Suboptimal

QR code

presentation

xms