1. **ForSMedia** is an advanced platform for social media data analysis for customer profiles enrichment. The important differentiator of the solution is its ability to process very large number of customer profiles automatically to find new information about customer interests, hobby, favorite music etc.

2. The motivation of the project was rather obvious. The information about clients is very important for companies in all industries. The more we know about people we are dealing with the more success we can achieve. Traditionally main sources of knowledge about clients were internal systems of organization. Now when SN are becoming more and more popular a lot of interesting data can be discovered there. New big data technologies allow extract this data and convert it to information valuable for business.

3. What SM can tell us about people? First of all it is attributes explicitly indicated by SN users in their profile. But not only this – important facts are contained in SN implicitly. Reading posts, descriptions and other texts we can discover interesting facts about hobbies, opinions, favorite movies and so on. It is very important to get not only explicit but also implicit information

4. Taking into account this motivation we created ForsMedia with the following functionality:
   - Social network data acquisition
   - Customer identification in social networks
   - Linguistic processing and text enrichment from posts, subscription groups, comments
   - Revealing new customer attributes by data mining technics
   - Merging social networks user data into unified customer profile
   - Analysis and Discovery

Let's look at each function in more details

5. Collecting data from SN. We are collecting information from Social Networks, pre-processing it and store in our Hadoop cluster. It allows us to find information about clients within reasonable time. ForSMedia provides several methods to collect and monitor SN data. The first method is to use public API provided by any popular Social Network. It is very easy to integrate API into any application and this is the advantage of the method. But the quality of API can be very low. Some of SN API are very open and allow to get a lot of different data but some of them gives almost nothing. It is always possible to parse web pages, extract data and then monitor it using crawlers. This method is very labor & time consuming, but here we doesn't depend on SN owner. And the third method is to buy data from a company which collect and update all data. The main advantage – high speed of data acquisition but it may expensive. The most reasonable way to collect and monitor data is to combine all three methods depending on the requirements of a particular project.

6. Customer Identification in SN. Customer identification in SN means the detection of all user profiles in SN that correspond the customer data provided for identification. The initial data for identification may be first name, last name, date of birth, city. Usually it is enough to find profiles but of course additional information like company, address, phone number can narrow the search. It is important to take into account that SN data are incomplete, not standard and sometime invalid. So short name or nick name can be indicated instead of full name, and some users of SN specify only day & month of birth without the year. It means that the user profile can correspondent to a given customer only with some confidence. ForsMedia supports confidence identification level based on our original algorithms of data cleansing& normalization and score calculation.

7. Getting more information about SN users. After detection of user profiles which correspondent to each customer a number of attributes in the user profile are available. ForSMedia provide standardization & cleansing of names, cities, addresses and so on. It is important to mention that valuable information about SN user is contained implicitly in posts, comments, description of subscription group and other texts. ForsMedia uses special linquistics & statistics technics to extract interests, favorite movies, etc from these texts.

8. The results of text processing can be not accurate due to the ambiguity of natural language and statistical assumptions that are not always true in real life. In ForSMedia approach is the joint usage of both technics than improve the quality and accuracy of results. But in all cases a level of confidence is provided for any fact extracted from text.

9. Creating unified customer profile. There may be several SN users that correspond to the same customer with high confidence or identification. In this case these users should be merged into unified customer profile. It can be done in different ways depending on the requirements and tasks. For example constructing the list of interests we can select only interests presented in every matched profile or it is possible to unite all interests of all matched users into one extended list of interests. In fact the merging algorithm can be more complicated than simple unions and intersections. We can use also confident rates resulted from linguistic processing.

10. Analysis of data is based on Data Discovery technology that allows to use intuitive search and discovery in contrast with traditional Business Intelligence. As a tool we are using Oracle Bid Data Discovery.

11. ForSMedia is based on Open source software. All data are stored in Hadoop cluster with the usage of Hbase and Hive and processing is implemented with R, Python and fact extraction software from Russian linguistic company RCO. ForSMedia can be installed on any hardware platform that meets the requirements. We have also the version of ForSMedia which is certified on a special software-hardware platform – Oracle Big Data

Appliance and Oracle Exalytic to deliver speed, reliability and scaleability. The first machine is Hadoop Machine and the second one is Oracle Analytics Machine.

About FORS. FORS was etablished in 1991. All this years the company offers full range of IT-services, including IT-consulting, development, integration, implementation and maintenance of systems and applications. FORS is one of the largest developers of Oracle-based systems in Europe delivering Oracle optimized infrastructure software-hardware solutions. FORS is more than 500 world-class specialists, over 1000 successfully implemented IT projects for finance, telco, government, medicine, sports and other industries. FORS was organized by people from Russian Academy of Sciences, so research in a very important part of our activities. And we also participate in educational area. FORS has Education Center delivering a lot of certified IT courses and we are closely connected with Higher School of Economics particularly with the Business Informatic Department.

FORS has been dealing with Analytics for 20 years and is very experienced developer of DW and BI applications in all kind of industries. So it was absolutely natural to go into Big Data area. 5 years ago a special department was organized to accumulate Big Data Expertise and Knowledge. Of course we are rather close to Oracle Big Data products and are working together with Oracle at some accounts but we are using a lot of open source big data technologies. Here is our practical experience. I would pay a special attention to the project in Alpha bank where we have developed solution based on Oracle Big Data Appl. (Hadoop machine), very interesting project for Real Estate. For Social Media data analysis we created a special software platform ForsMedia which is the main topic of my presentation.