

الأكاديمية العربية الدولية



الأكاديمية العربية الدولية
Arab International Academy

الأكاديمية العربية الدولية المقررات الجامعية



بسم الله الرحمن الرحيم

جامعة المستقبل
علوم الإدارة وتكنولوجيا المعلومات

Data mining

إعداد المهندس/نبيل محمد لطف مصلي

Email:- nabil299@Gmail.com

مقدمة عن علم التنقيب عن البيانات

علم التنقيب عن البيانات Data Mining علم مستقل بحد ذاته ، ولو أن بعض الباحثين يرى ربطه بعلم استجلاب المعلومات Information Retrieval ، نظرا لأن هذا الأخير يعتبر أقدم بكثير من الـ Data Mining و لكن يرى العديد من الباحثين أيضا أن نضوج مواضيع هذا العلم يجعله علم مستقل بحد ذاته .

بعبارة بسيطة جدا (قد تكون مخلة بعض الشيء ولكن لغرض التبسيط سأسردها) يعتبر علم تنقيب البيانات العلم المسئول عن أساليب و طرق إنتاج المعلومات و قواعد المعرفة من خلال كم كبير من البيانات التي يتم التنقيب فيها و ربطها بأساليب علمية لنخرج بمعلومة أو معرفة جديدة

على سبيل المثال لا الحصر ، التصنيف الأوتوماتيكي للمستندات يعتبر من الجوانب التي يتناولها علم الـ Data Mining ، فهناك العديد من الخوارزميات التي من شأنها تصنيف المستندات بعد إعطائها كم هائل من البيانات مربوطة بنوع هذه البيانات و تدريبها على ذلك ثم إنشاء نظام متكامل للتصنيف اعتمادا على عملية التدريب

دعني أضرب مثال عملي أكثر ... الجميع يعرف فلاتر البريد المزعج (Spam Email) التي تستخدمها شركات مثل Yahoo! و Gmail و Hotmail ... الكل يعلم فائدتها و استخداماتها ، ولكن طريقة عملها يجهلها الكثير ... تعمل هذه البرمجيات باستخدام خوارزمية تعرف بـ Naive Bayesian ، هذه الخوارزمية أو القانون من قوانين علم الإحصاء ... لست هنا بصدد شرح هذه الخوارزمية ، ولكن لأذكر مثلا على استخدامات علم الـ Data Mining في التطبيقات ، حيث أصبح الـ Naive Bayesian Classifier يدرس بكثافة في تطبيقات هذا العلم .

ما هي أو ما معنى التنقيب في البيانات ؟

هي عملية بحث محوسب ويدوي عن معرفة من البيانات دون فرضيات مسبقة عما يمكن أن تكون هذه المعرفة.

تعريف آخر:-

هي عبارة عن تقنية تستخدم في عملية استخراج البيانات من قواعد بيانات عملت مثل قاعدة :

Data ware haws

Working with data mining

Dm : تمكّنك من الوصول إلي المعلومات التي تحتاجها لصناعة القرارات الذكية وذلك حول مشاكل الأعمال التجارية الصعبة .

هناك هدفان أساسيان للتنقيب في البيانات هم:

التنقيب الاستشراقي والتنقيب الوصفي. التنقيب الاستشراقي ينتج عنه نموذج عن النظام الذي تصفه البيانات المستخدمة في التنقيب. أما التنقيب الوصفي فنتج معلومات جديدة بناء على المعلومات الموجودة داخل البيانات المستخدمة في عملية التنقيب. العمليات الأساسية التي تعتبر ضمن عمليات أو مهام التنقيب هي:

١. التصنيف
٢. الانحدار
٣. العنقدة (التجميع)
٤. التلخيص
٥. نماذج الاعتماد
٦. الكشف عن التغيرات والانحرافات

هناك ثلاثة أهداف للتنقيب في البيانات :

- ١- من أجل تعليل بعض الظواهر المرئية. مثال: لماذا زادت نسبة المدخنين في الوطن العربي؟
- ٢- من أجل التثبت من نظرية ما. مثال: التثبت من النظرية التي تقول بأن الأسر الكبيرة تهتم بالضمان الصحي أكثر من الأسر الصغيرة عدداً.
- ٣- من أجل تحليل البيانات للحصول على علاقات جديدة وغير متوقعة. مثال: كيف سيكون الإنفاق العام إن كان ملازماً لعمليات خداع واسعة من قبل البطاقات الائتمانية.

وسائل التنقيب في البيانات :

هناك عدة وسائل مختلفة من أجل التنقيب في البيانات. اختيار الوسيلة المناسبة يعتمد على طبيعة البيانات تحت الدراسة وعلى حجمها. يمكن إجراء عملية التنقيب في البيانات بالمقارنة مع سوق البيانات ومخزن البيانات.

بعض من هذه الوسائل هي :

- التفكير واستخلاص النتائج والقوانين من أمثلة حية Case-Based reasoning
- الكشف عن قانون Rule Discovery: البحث عن منوال معين أو علاقة معينة في جزئية كبيرة من البيانات
- معالجة الإشارات Signal Processing: إيجاد الظواهر المتشابهة مع بعضها البعض
- شبكات نووية: Neural Nets تطوير نماذج قابلة للتنبؤ بالنتائج. هذه النماذج تم تطويرها بناءً على أسس تم استنباطها من عقل الإنسان.
- منحنيات غير ثابتة: Fractals تصغير البيانات الكبيرة من دون ضياع المعلومات

تطبيقات التنقيب في البيانات :

وسائل التنقيب في البيانات تُستعمل وينجح في الكثير من التطبيقات الحقيقية حول العالم. التطبيقات التالية تشمل بعضاً من الأمثلة

- كتابة تقرير مختصر عن فئة معينة: Profiling Populations تطوير وإنشاء تقارير موجزة عن الزبائن المهمين وعن بطاقات الائتمان.
- تحليل النزعة التجارية: Analysis of Business Trend إيجاد الأسواق ذات قدرات النمو القوية أو الضعيفة.
- التسويق لفئة معينة: Target Marketing إيجاد الزبائن من أجل منح التخفيضات لهم لسبب معين.

- تحليل الاستعمال Usage Analysis: إيجاد منوال معين لاستعمال الخدمات والسلع

- فعالية الحملة Campaign Effectiveness: مقارنة استراتيجيات الحملات مع بعضها البعض من أجل إيجاد أكثرها فعالية وتأثيراً.

- جاذبية السلعة: إيجاد السلع التي تباع مع بعضها البعض.

تطبيقات التنقيب في البيانات بدأت تنمو بصورة كبيرة للأسباب التالية:

1- كمية البيانات الموجودة في مخزن البيانات وسوق البيانات تنمو بصورة أساسية (exponential) ومن أجل ذلك، فإن المستخدم يحتاج إلى أدوات متطورة من مثل التنقيب في البيانات من أجل استخلاص الفائدة والمعرفة من هذه البيانات.

2- الكثير من أدوات التنقيب عن البيانات بدأت تظهر مؤخراً، وكل أداة أفضل من الأخرى.

3- المنافسة الشديدة الموجودة في السوق تدفع الشركات إلى الاستفادة القصوى من البيانات التي بيدها. عمليات التنقيب في البيانات تفعل ذلك تماماً.

Microsoft SQL Server Analysis (SSAS) Microsoft SQL Server

يقوم بتزويد الأدوات للـ Dm والتي تقوم بتعريف القواعد والنماذج في بياناتك ولذلك يمكن تحديد لماذا الأشياء تحدث والتنبيه بما سوف يحدث في المستقبل .

Creating an analysis services project :

إنشاء مشاريع خدمات التحليل : لإنشاء للـ Dm : أولاً أن نقوم بإنشاء analysis services project وحيث إذا نقوم بتشكيل أو إعداد لمصادر البيانات وعرض مصدر البيانات وذلك للمشروع . مصدر البيانات يقوم بتعريف مصدر الاتصال ومعلومات التحويل التي تتصل بمصدر البيانات الموجودة في نموذج ال Dm الأساسية :

Adding mining structures إضافة هياكل

Analysis server project تنقيب لمشروع خدمات التحليل .
بعد أن تقوم بإضافة Analysis server project يمكنك أن تقوم بعملية إضافة هياكل تنقيبيه إليها وكذلك واحدة أو أكثر من النماذج التنقيبية التي تعتمد منها على الأخرى .

الهياكل التنقيبية :

تشمل على جداول وأعمدة وتكون مشتقة من مصادر بيانات عرض متوفراً ومن عملية OLAP (: OLAP Online Fanatical process) أي عملية التحويل المباشر من موقع الانترنت .

Working with Dm models :

قبل أن نستطيع استخدام نماذج التنقيب التي تقوم أنت بتعريفها يجب عليك أن تقوم بعملية المعالجة لها وذلك الخدمات التحليل analysis services

تستطيع المرور بعملية تدريب البيانات من خلال الخوارزميات لإنشاء النتبؤات : creating predictions

- الهدف الرئيسي من معظم مشاريع تنقيب البيانات يكون مستخدماً لنماذج التنقيب وذلك لإنشاء التنبؤات بعد أن يقوم بعملية استكشاف ومقارنة نماذج التنقيب .
- تستطيع أن تستخدم واحدة لإنشاء نماذج التنبؤات

خدمات التحليل :

- تقوم بتزويدك بلغة الاستعلامات Dmx = Data miming Extension
- عرف كلاً من المصطلحات التالية (Dmx – ssAsd) التي تستخدم لعملية إنشاء التنبؤات
- SQL server Reporting :
- بعد أن تقوم بإنشاء نماذج التنقيب بإمكانك أن توزع النتائج على مستوى أوسع ، ويمكنك استخدام التقارير الموجودة في (SSAS) لإنشاء التقارير والتي يمكن أن تقوم بتمثيل المعلومات المحتوية عليها نماذج التنقيب كذلك يمكن استخدام استعلام (DMX) كأقصى للتقارير .
- وتستطيع أن تأخذها كميزة لعملية الإثبات وتهيئة المميزات والتي تكون في خدمات التقارير
- Working programinatically unit Dmx
- العمل برمجي مع Dmx
- الـ Dmx تقوم بتزويدك بالجمال البرمجية التي تستطيع أن تستخدمها لإنشاء وتدريب نماذج تنقيب البيانات كذلك بإمكانك أداء عدة مهمات باستخدام إتحاد كلاً من Xml لعملية التحليل (ASSL) أو باستخدام Analysis services languages
- Analysis management object = Amo تحليل كائنات إدارية .
- تشغيل حزمة SQL :
- لكي يصبح SQL server شغال بشكل اتوماتيكي كلما أفلح الحاسوب تقوم بالآتي :
- ١. نقوم بالدخول الي Servies عن طريق ادوات النظام administrator tool
- ٢. نقوم بتحديد SQL server تم نقوم بالدخول خصائص
- ٣. من القائمة start up type نقوم باختيار الخيار أوتوماتيك (Automatic) .
- ٤. ثم نقوم بتحديد SQL server Agentn من services - فائدة يقوم بتشغيل التقارير والاستعلامات ثم ندخل إلي خصائصها ونعملها : أوتوماتيك .
- تفعيل الخدمات الأخرى التابعة لـ SQL server : للعمل مع خدمات SQL server يجب تفعيلها قبل العمل معها وبالعالم يتم تفعيل هذه الخدمات لتعمل معها يدوياً بغرض تحلي وتقيب (استخراج) البيانات من قاعدة البيانات .
- ومن أهم هذه الخدمات الثانوية :
- SQL server Analysis services (SSAS) هذه الخدمة تعمل على عملية تحليل البيانات ويمكن انشاء فيها Structure model وها يعني انه يوجد مصدر بيانات Data Minig wizard (معالجة تنقيب البيانات)
- طريقة دخول الـ SQL server Analysis services (SSAS) :
- ١. نقوم بتشغيل البرنامج SQL server Manegement studio .
- ٢. نقوم بعملية الاتصال بقاعدة البيانات والـ server
- ملاحظة (يمكن تشغيل أكثر من قاعدة عن طريق سثقرثق وذلك عن طريق الزر options ثم نقوم باختيار قاعدة بيانات الافتراضية)
- وتتم عملية الاتصال لـ server من النوع Analysis service
- ١. من القائمة server type نقوم باختيار اسم الخادم المناسب المفعّل في الحاسوب وعادة يكون اسم السيرفر المحلي هو نفس اسم حاسوبك .

ويمكنك الدخول الي خوادم أخرى موجودة في الشبكة من النوع lan ، wan وذلك كالتالي :

- من القائمة server name نختار التوبيو network service
- في حالة وجود شبكة من النوع lan فقط فلا يمكن الدخول فقط الي الاجهزة الموجودة في نفس المعمل أو نفس المبني .

* سؤال عملي (قم بالدخول الي اسم الخادم المسمى pc3 الموجود في الشبكة وذلك عن طريق خادم من النوع Analysis server

الإجراءات التي يجب تطبيقها على قواعد البيانات قبل عملية التنقيب أو قبل عملية تطبيقات data minig

الأجراء TakeOffLine

عدم تمكين الخوادم الاخرى الي الدخول إلى نفس هذه القاعدة حيث بعد هذا الإجراء سوف تكون عمل بشكل حقل .

بالضغط على زر الماوس الايمن في قاعدة البيانات ونضغط Tasks ثم TakeOffLine الاجراء الثاني Copy database : عملية النسخ نقوم بعملية نسخ قاعدة البيانات من instance الي instAnCE آخر وذلك لغرض حماية البيانات وتبقي كقاعدة حيث اذا ما تضررت القاعدة الحالية لاجراء عملية copy database وتتم عملية اجراءات النسخ لقاعدة البيانات .

١. نقوم بتحديد قاعدة البيانات المراد قفلها إلى instance آخر واختر copy data
 ٢. نقوم بضغط next
 ٣. نقوم بتحديد اسم الخادم المصدر source
 ٤. ثم نقوم بتحديد نوع التحويل ثم next
 ٥. نقوم بتحديد الخادم للهدف المراد نقل قاعدة البيانات الموجود المراد نقل قاعدة البيانات اليها المحتوى على instance
 ٦. نقوم بالضغط على next ثم نقوم باختيار الطريقة التي سوف نقوم باستخدامه في عملية النسخ والنقل .
 - أ- (أ) الطريقة الأولى use the attach and attach method : هذه الطريقة سريعة ولاكنها تشترط أن تكون قاعدة البيانات على شكل offhinc
 - ب- الطريقة الثانية use the SQL Management object method : هذه الطريقة بطيئة ولاكنها تعمل في وان كانت قاعدة البيانات on line
- لكي تجعل قاعدة البيانات تقوم بعملية الاتصال من جديد بالـ instance اذا ما فشلت عملية النسخ فإننا نقوم بتنفيذ الخيار السابق ،
- ويقصد بعملية reattach عملية اتصال قاعدة ما موجودة في مسار ما إلى instance موجود في مسار ما إلى instance موجودا في الخادم العام .
٧. نقوم بالضغط على زر next ثم نقوم بتحديد قاعدة البيانات .
- ملاحظة (بحالة وجود مشكلة في عملية النقل في النموذج أعلاه نرجع الي قاعدة البيانات ونجعلها online وبعد ذلك نرجع إلى مربع الحوار أعلاه ثم نختار الطريقة الثانية) .
٨. نقوم بكتابة اسم قاعدة البيانات الجديدة وإذا اردت استبدالها بقاعدة بيانات سابقة موجوده نسميها بنفس الاسم لكي يحدث التبديل ثم next
- ملاحظة (لكي تنجح العملية كما يجب أن يكون SQL server Agent مفعّل .
- عملية الـ Attach لقاعدة البيانات :
- الغرض في هذه العملية هو عملية اتصال قاعدة البيانات في الخادم وهي حالياً توجد في أي مسار ، ودائماً نقوم بالعملية في قاعدة البيانات عندما نقوم بتحميلها لزبون .
- وذلك لكي تستطيع واجهه البيانات بالاتصال بها .

استخدام الأداة Analyze في عملية التحليل : وهي أداة من أدوات تنقيب البيانات :

- ١- تحدد قاعدة البيانات .
- ٢- من القائمة Tool نختار Database wizard Analyze
- ٣- ثم تظهر رسالة نختار الخيار الأول statistics والغرض من هذا الخيار هو تحليل البيانات الموجودة في قاموس البيانات وتخزينها في قاموس البيانات ثم الضغط على زر next
- ٤- يظهر مربع حوار فنقوم باختيار الكائنات المراد تحليلها ليتم نقلها الي الصفحة اليمنى وذلك عن طريق تحديدها والضغط على زر النقل <
- وعلي سبيل المثال سوف نقوم هنا باستخدام المستخدم scott وتحديد خيرا ونقلها الي الجهة اليمنى تساعد selcted object ثم الضغط على زر next
- تظهر رسالة يتم اختيار الجداول كاملة وفي كل واحد يتم تفعيل الخيار الاول يم تظهر رسالة نختار التالي ثم تظهر رسالة نتخار finish ثم تظهر رسالة نجاح العملية .

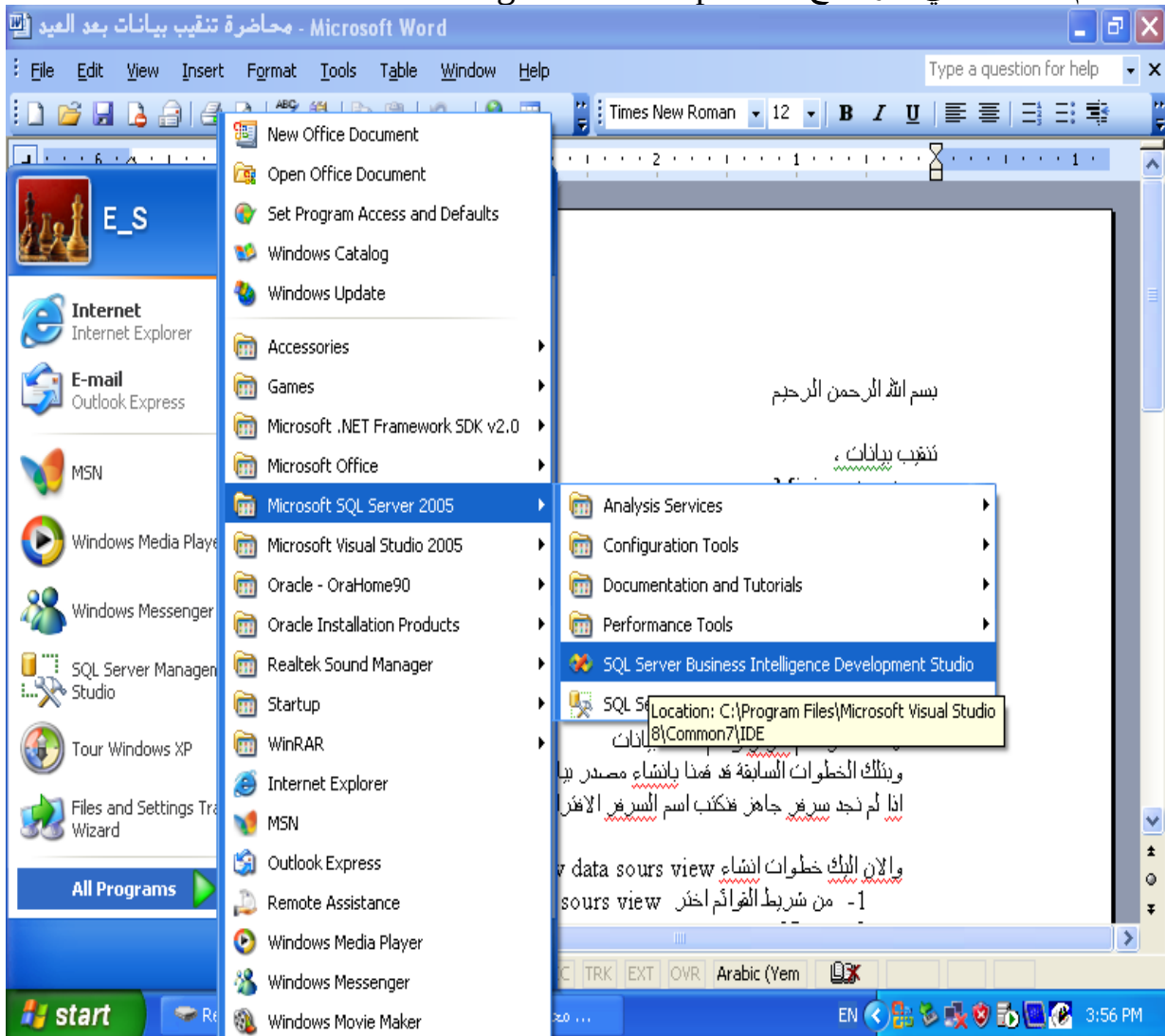
ناتي الآن الى الجانب العملي لتنقيب بيانات وسوف نستخدم للتطبيق Microsoft SQL Server 2005

Mining structures

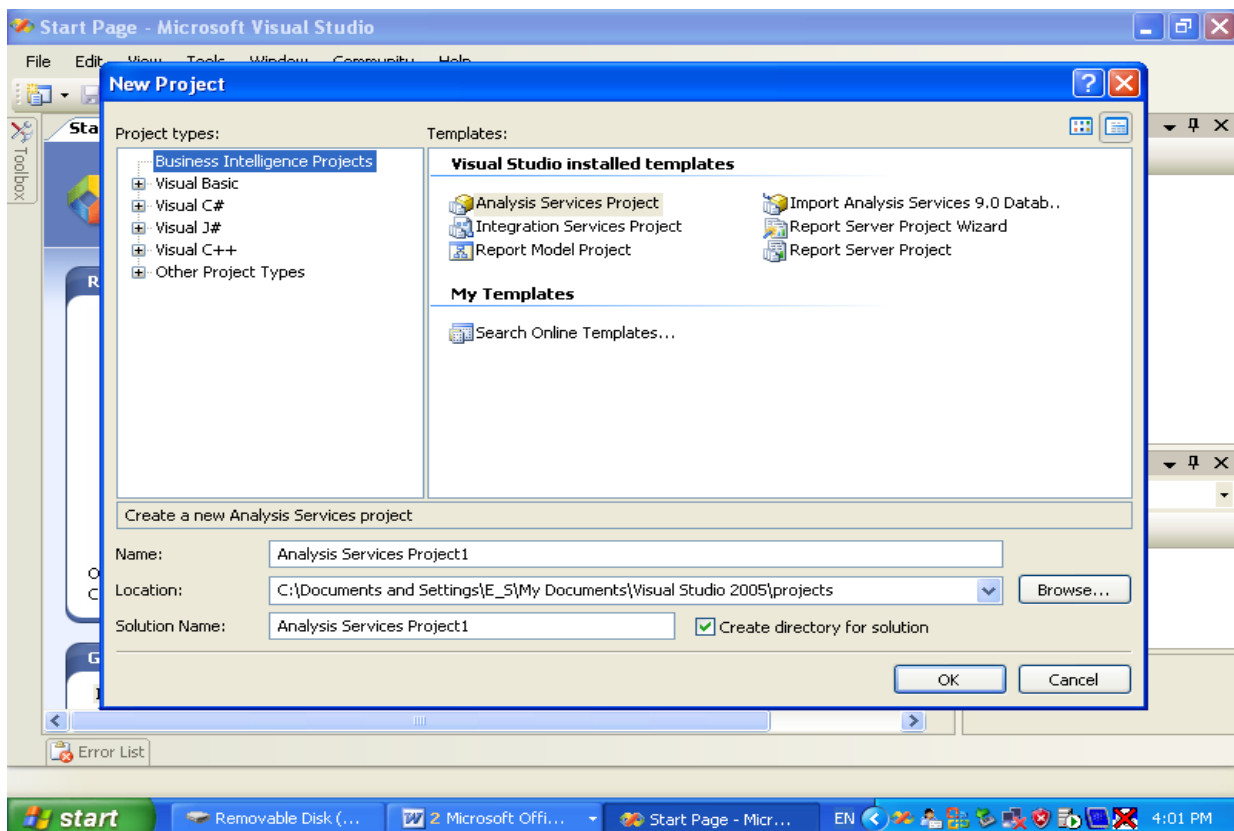
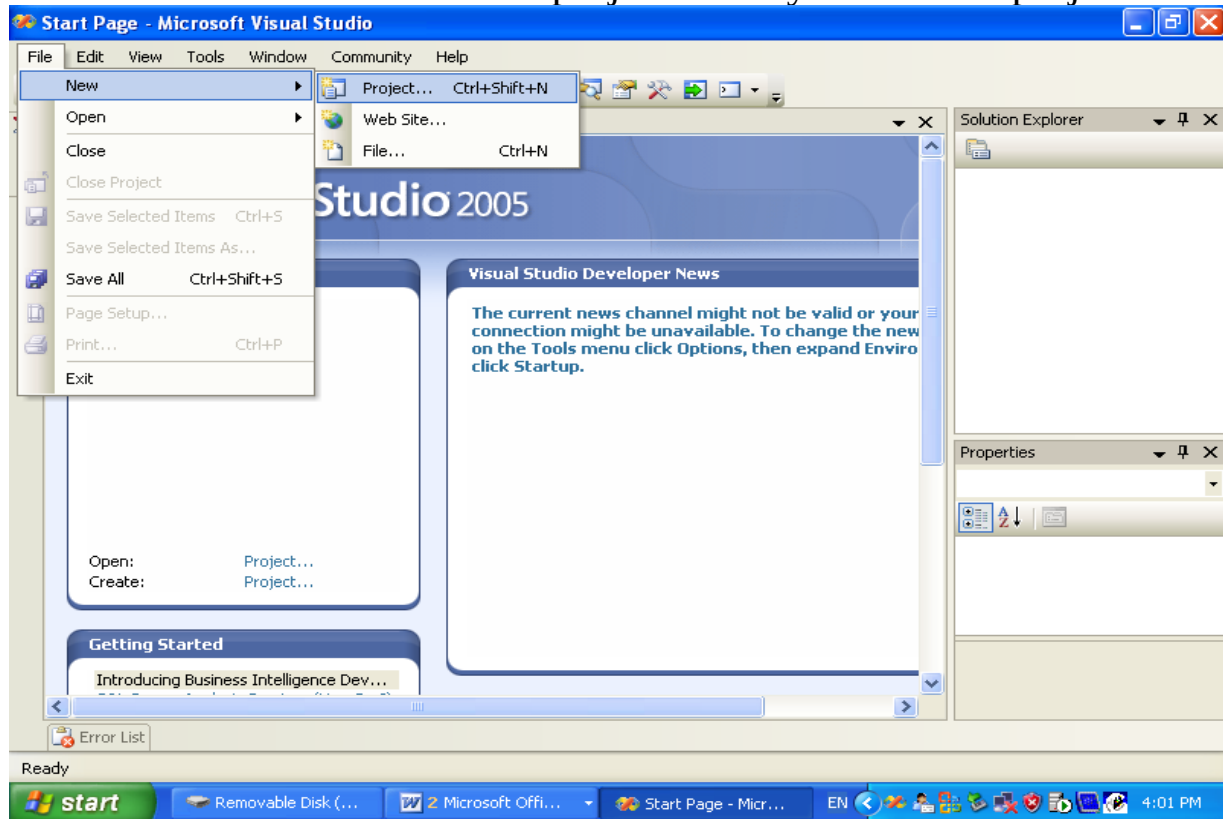
Mining models

الخطوات العملية لبناء هياكل التنقيب ونماذج التنقيب :

١- نقوم بالدخول الي البرنامج bussiness intelligence development

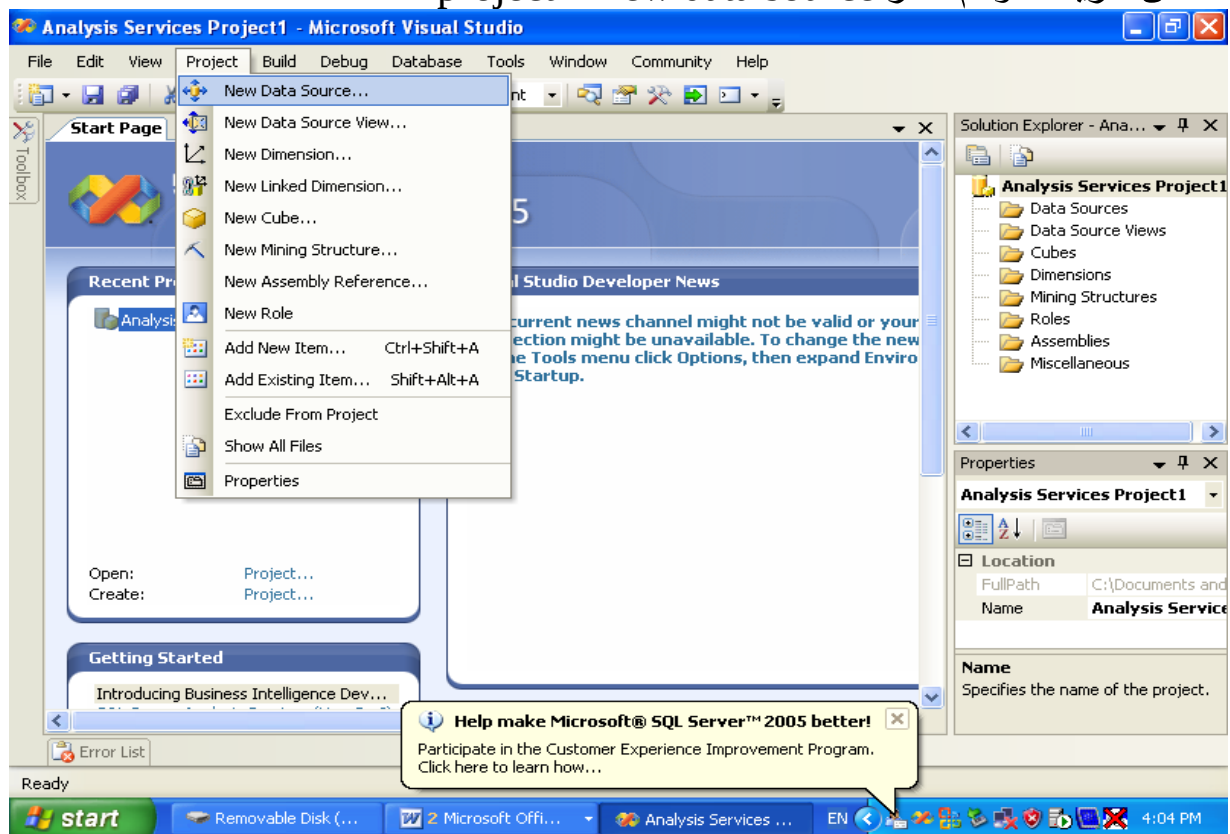


٢- من file + new project + analysis services project

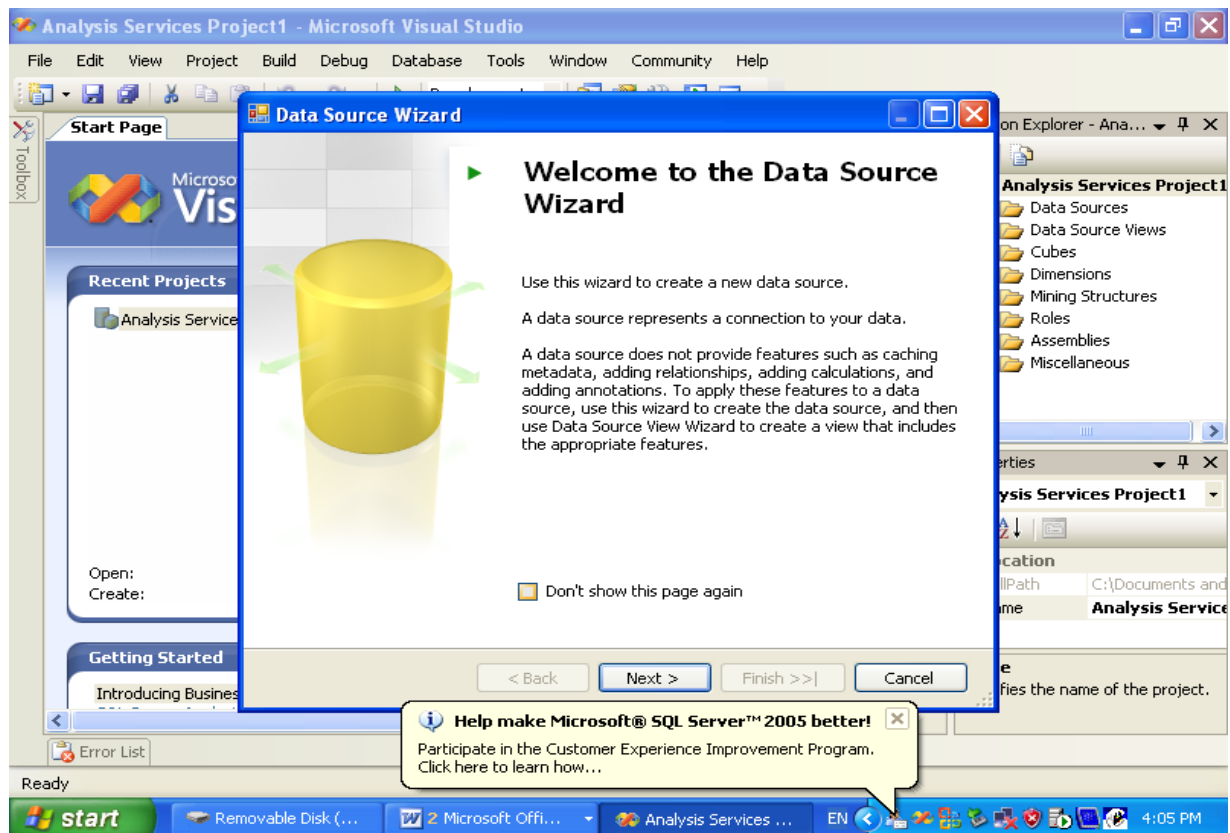


Ok -٣

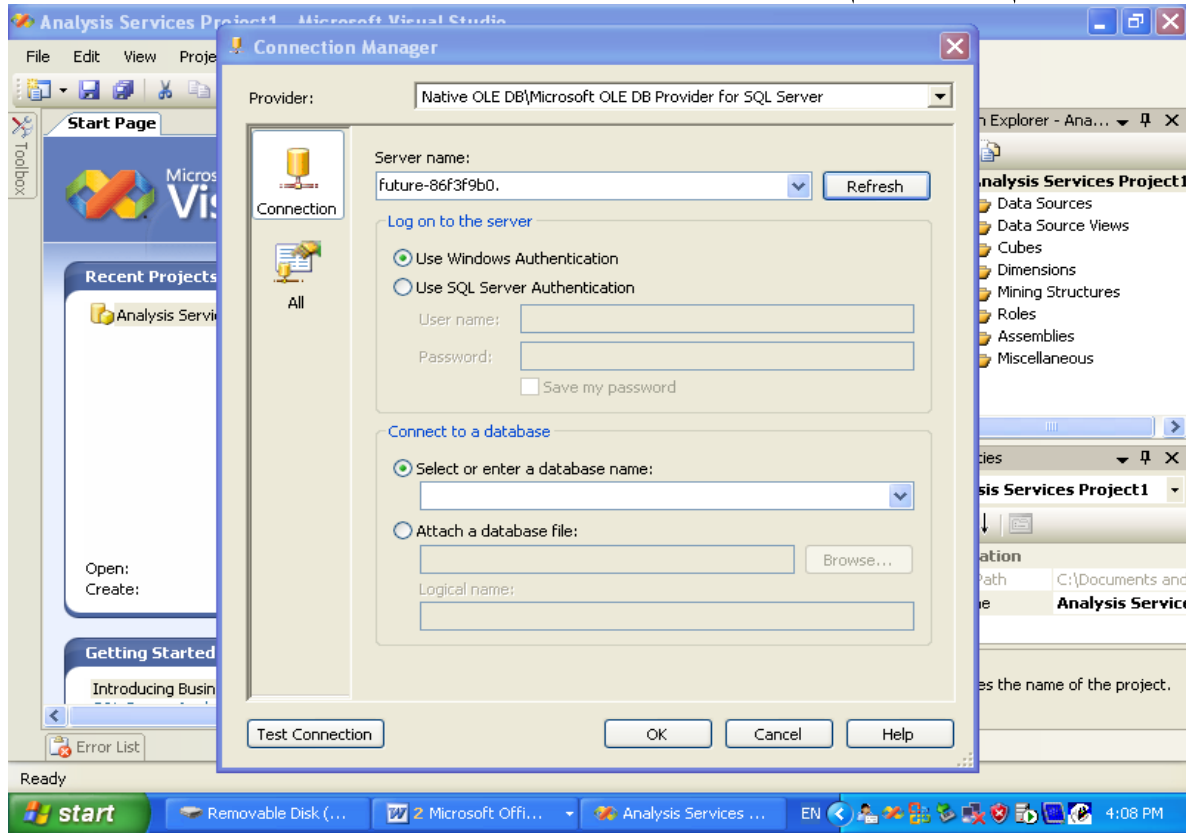
٤- من شريط القوائم اختر project + new data source



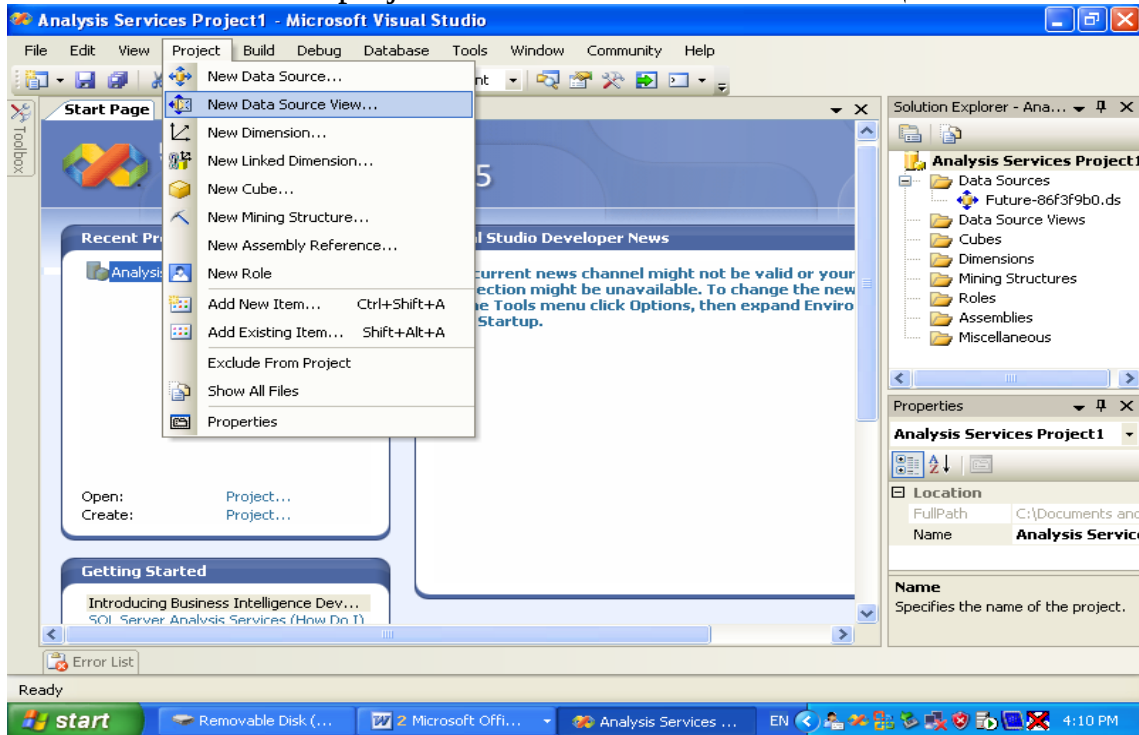
Next ->



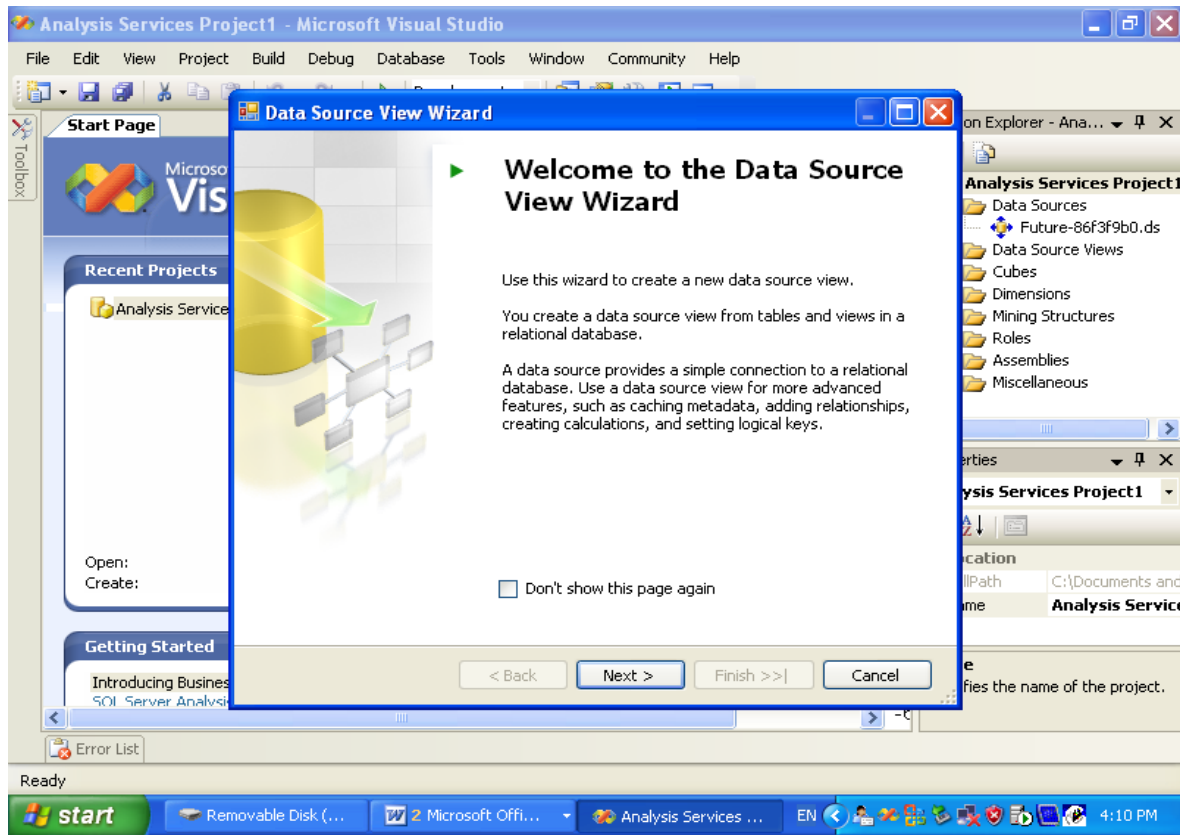
٦- نختار اسم سرفر واسم قاعدة بيانات



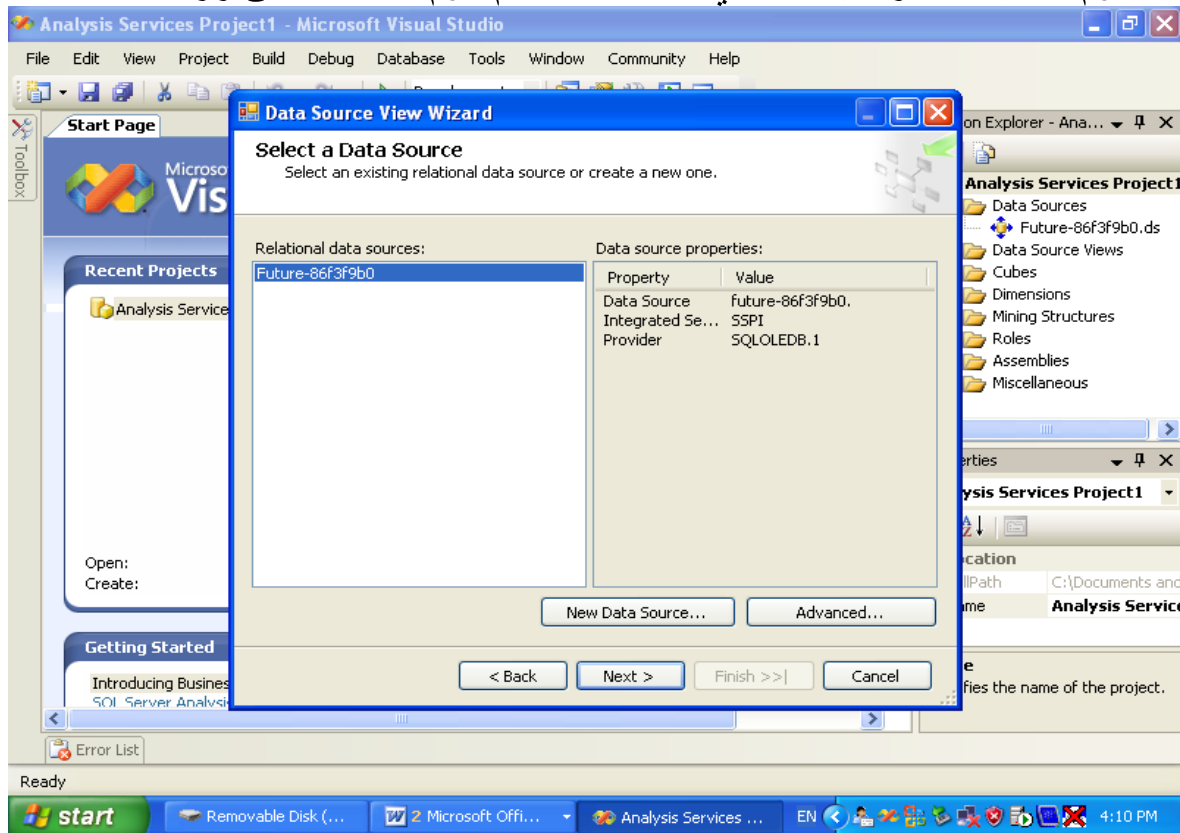
وبتلك الخطوات السابقة قد قمنا بانشاء مصدر بيانات جديد new data source
 اذا لم نجد سرفر جاهز فنكتب اسم السرفر الافتراضي وهو اللوكل هوست localhost
 والان اليك خطوات انشاء new data source view
 ١- من شريط القوائم اختر project + new data source view



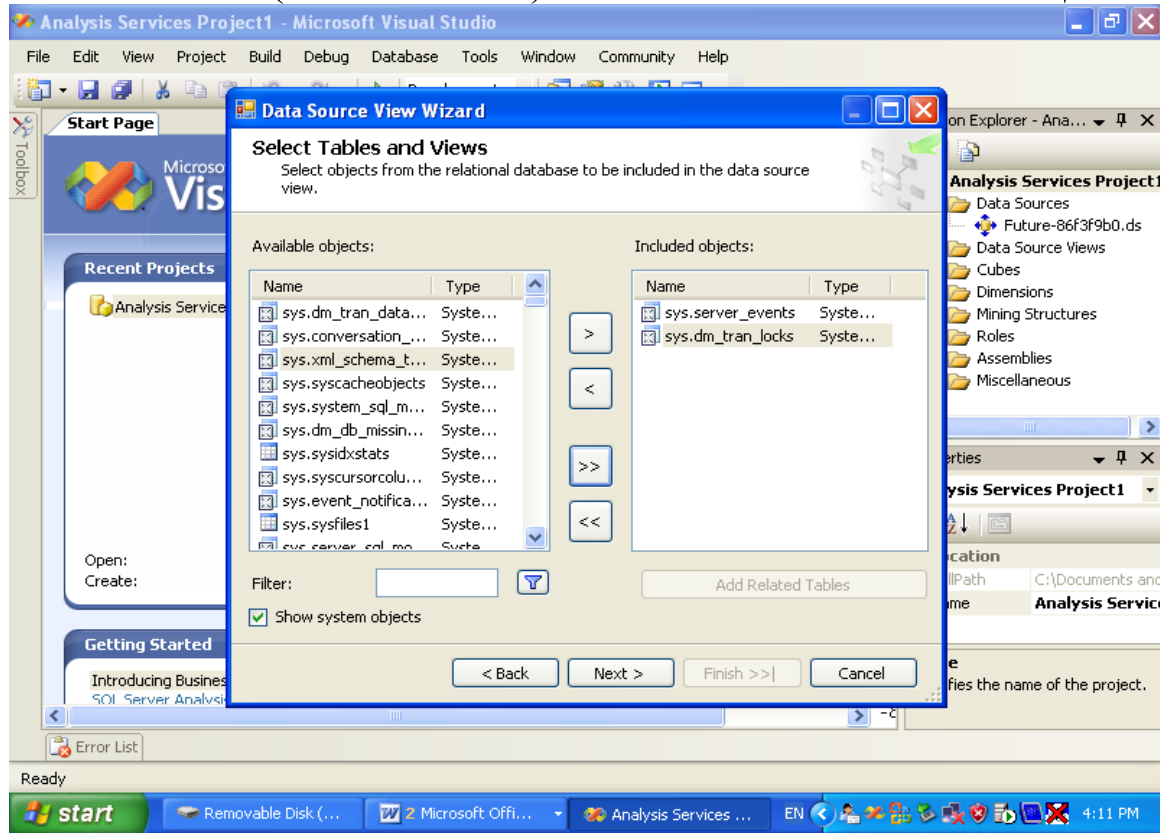
Next -٢



٣- نقوم بتحديد مصدر البيانات الذي انشأنه سابقا ثم نقوم بالضغط على زر next



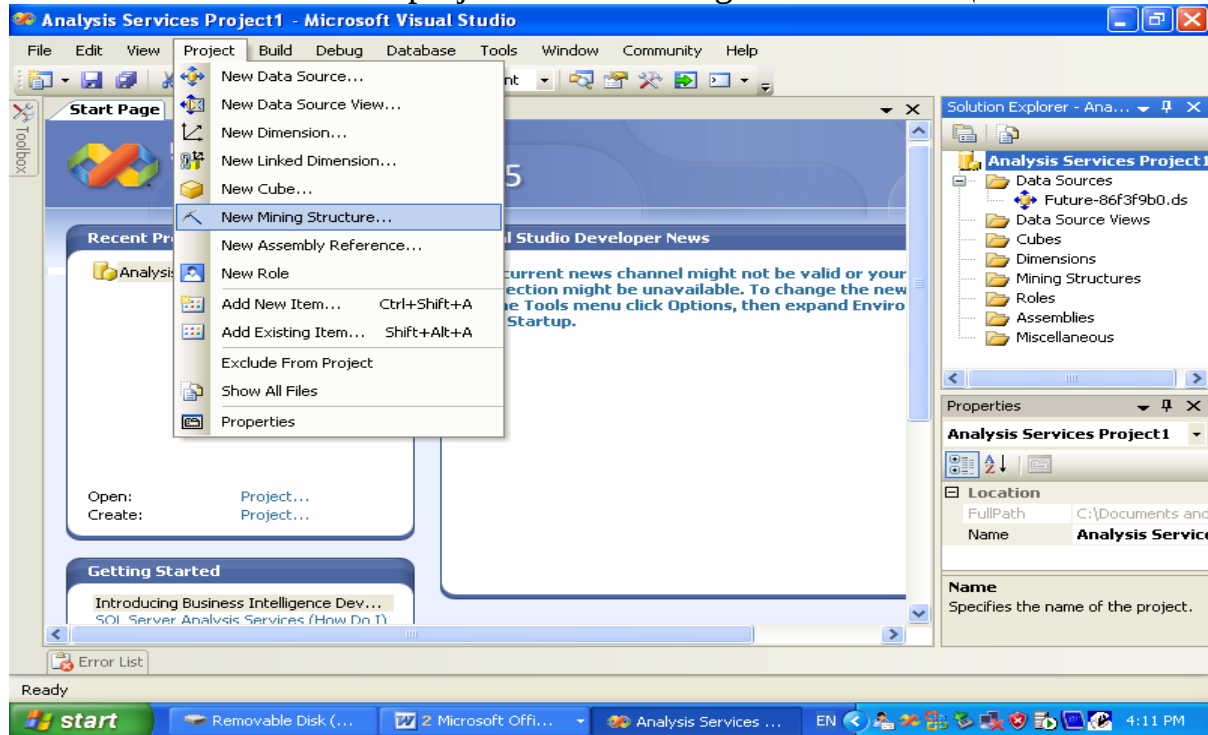
٤- نقوم بتحديد الجداول المراد بناء الهياكل لها (هياكل التنقيب لها)



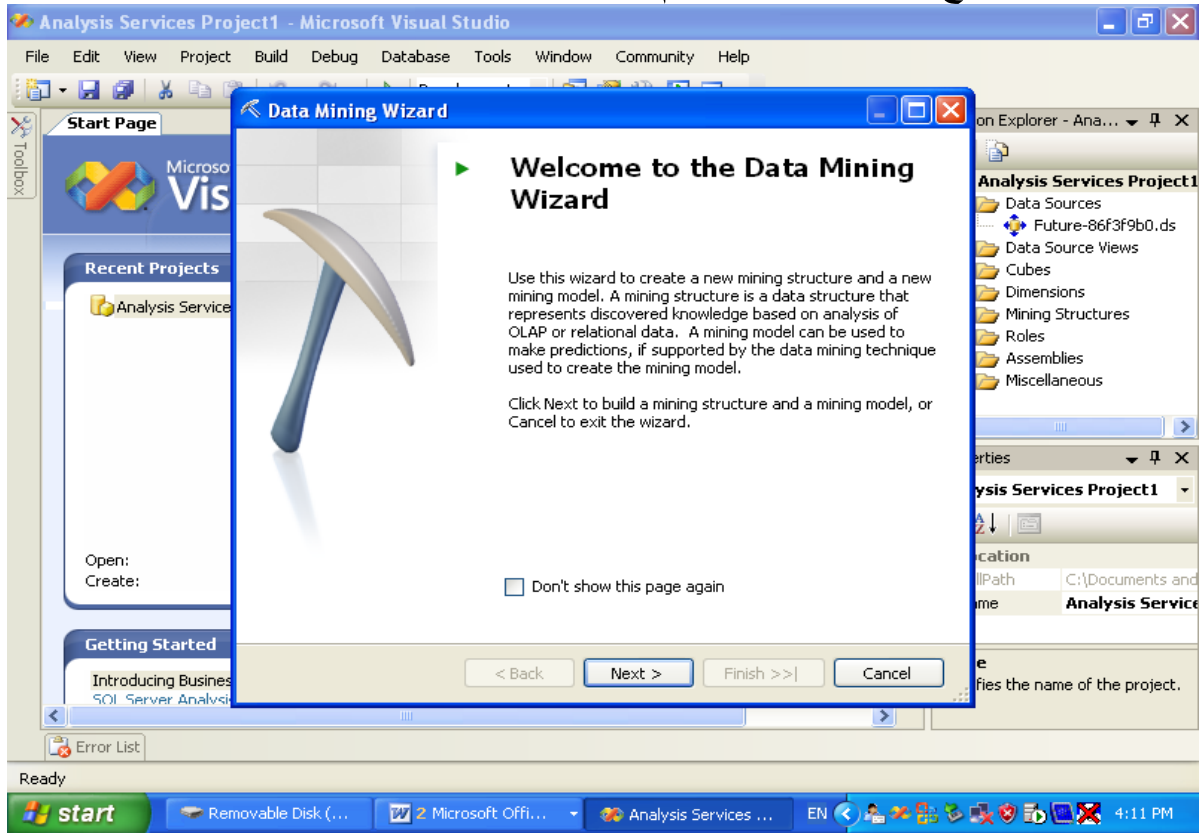
٥- ثم انقر على زر next + finish

الان نقوم بعملية انشاء new Mining structures وذلك بناء على data source وعلى view

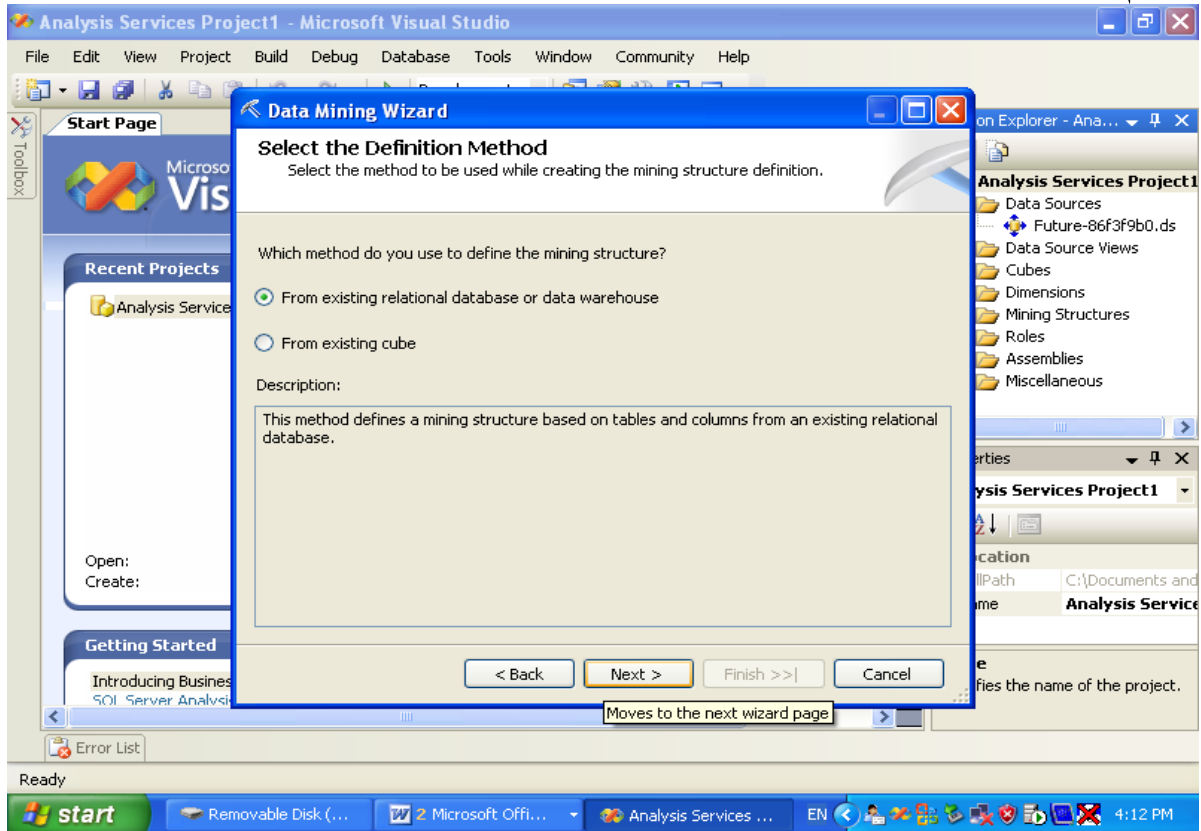
١- من شريط القوائم اختر project + new Mining structures



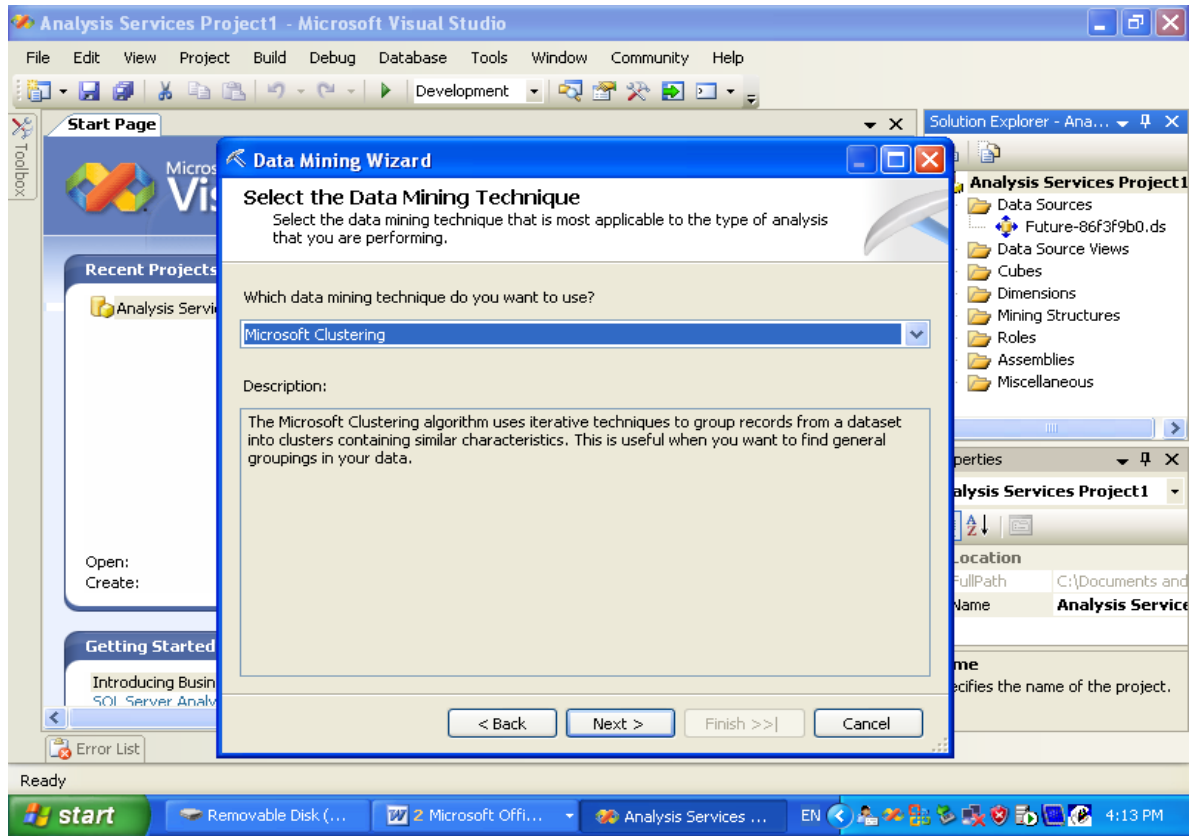
٢- يظهر لنا معالج تنقيب البيانات فنقوم بالضغط على الزر Next



٣- ثم النقر على الزر Next



٤- ومن (أي من تقنية تنقيب البيانات which data minig technique do you want use) تريد ان تستخدمها ؟) (ومن هنا نقوم باختيار Microsoft clustering لانها تجمع للبيانات



٥- ثم النقر على الزر next

٦- من قائمة input tables نقوم بتحديد وتفعيل علامة الـ case في الجداول المراد التعامل معها

٧- بعد النقر على زر التالي ثم اختر الحقول المراد التعامل معها ثم النقر على الزر next

٨- ثم النقر على زر finish

استخدام معالج تنقيب البيانات datat minig wizard لعملية انشاء minig structure هيكلية التنقيب و هيكلية النماذج minig model

بعد ان قمنا في المحاضرة السابقة بعملية انشاء new data source ثم إنشاء new data source view فأننا نستطيع في هذه المحاضرة أن نقوم بإنشاء ما ذكرناه وذلك بواسطة المعالج تنقيب البيانات وذلك كالآتي :

١- من قائمة project نقوم باختيار new minig structure

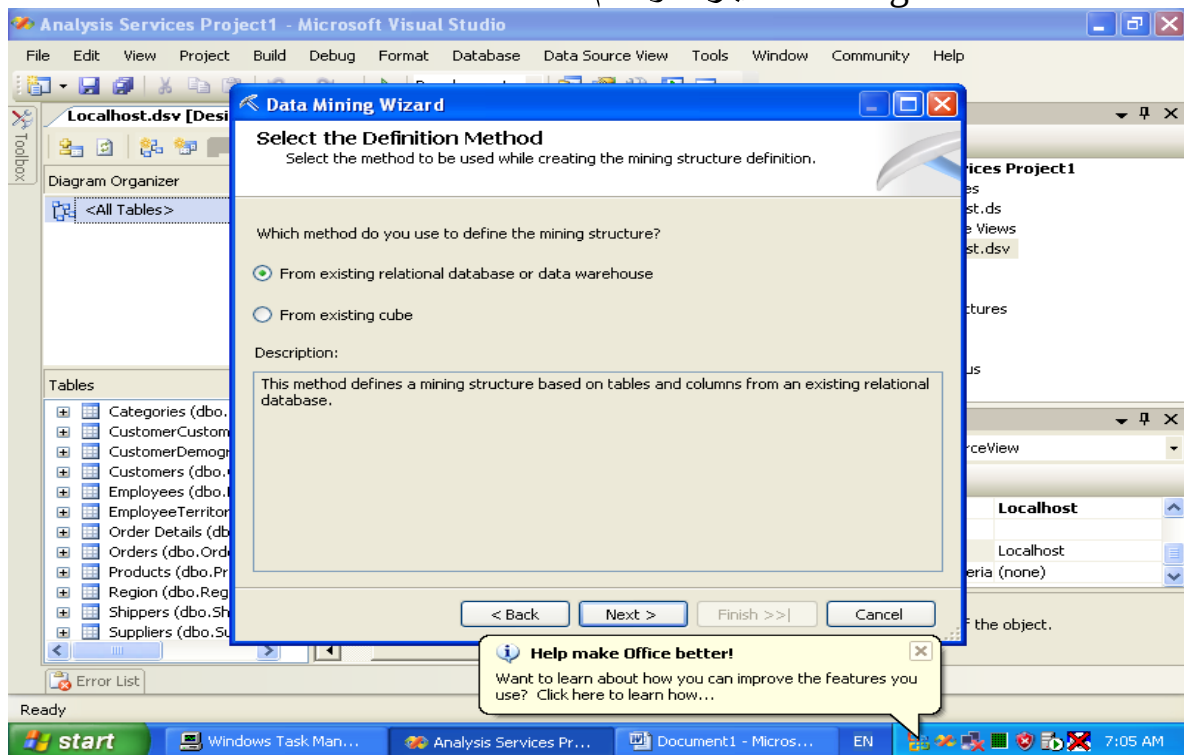
ملاحظة (-1 يقصد بها بأنها عبارة عن هياكل البيانات وتراكيب البيانات التي تمثل المعرفة المكتشفة المبنية على تحليل تقنية الـ olap او عملية بيانات علائقية .

ملاحظة (يقصد بعملية minig model بانها تقنية تستخدم لعملية التنبؤات المستقبلية ليتم اتخاذ القرارات المناسبة ويتم انشائها بواسطة data minig)



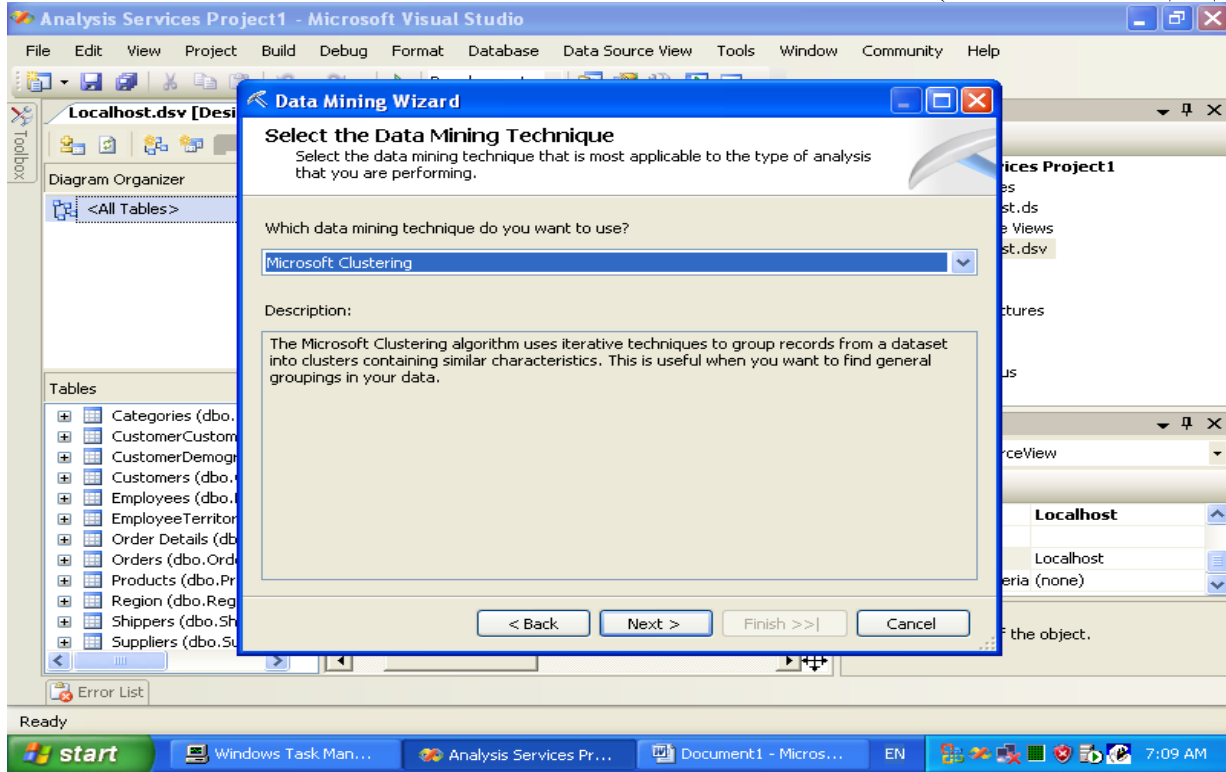
٢- انقر على الزر next

٣- فنجد معالج بيانات يخبرنا أي من الطريقة التي سوف نستخدمها لتعريب ويفضل في هذه الحالة minig structure الاختيار الاول ثم next

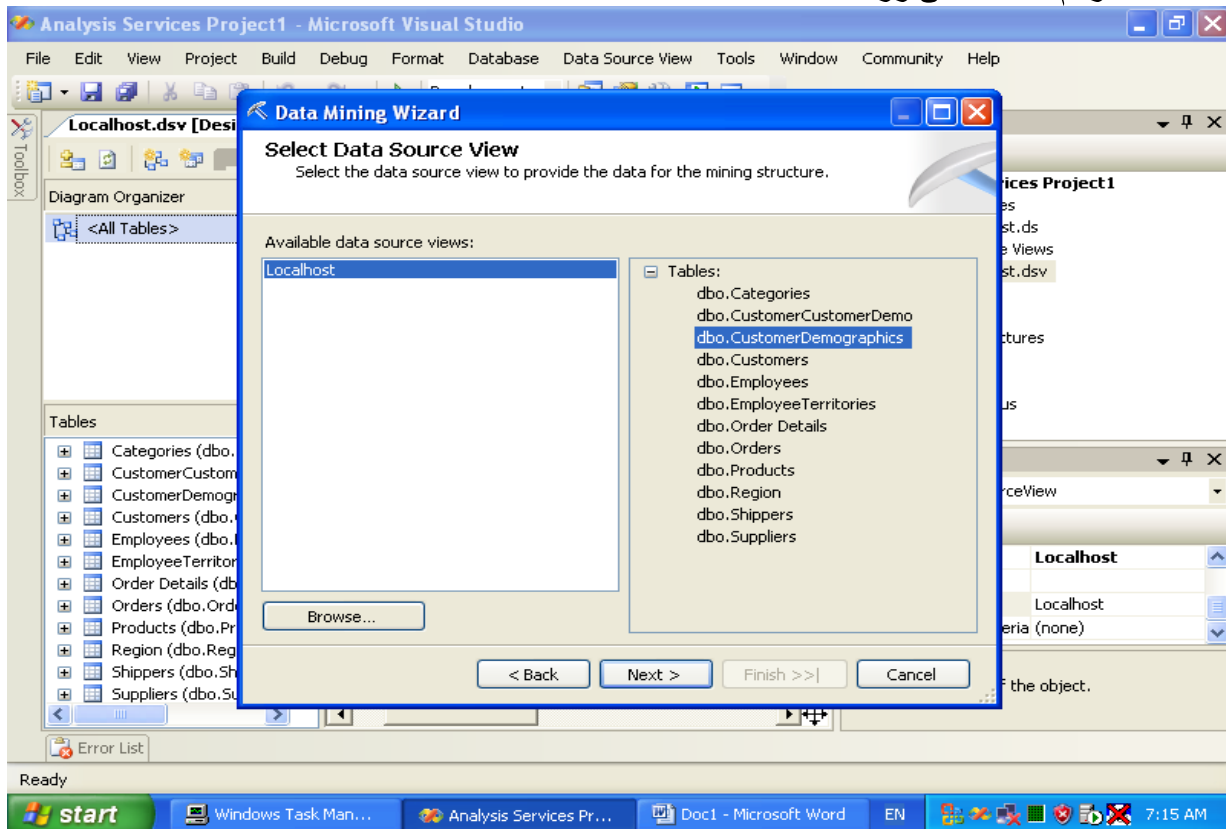


٤- نجد مربع حوار لمعالج التنقيب يخبرنا أي من تقنية تنقيب البيانات المراد استخدامها ويفضل في هذه الحالة استخدام تقنية microsoft clustering

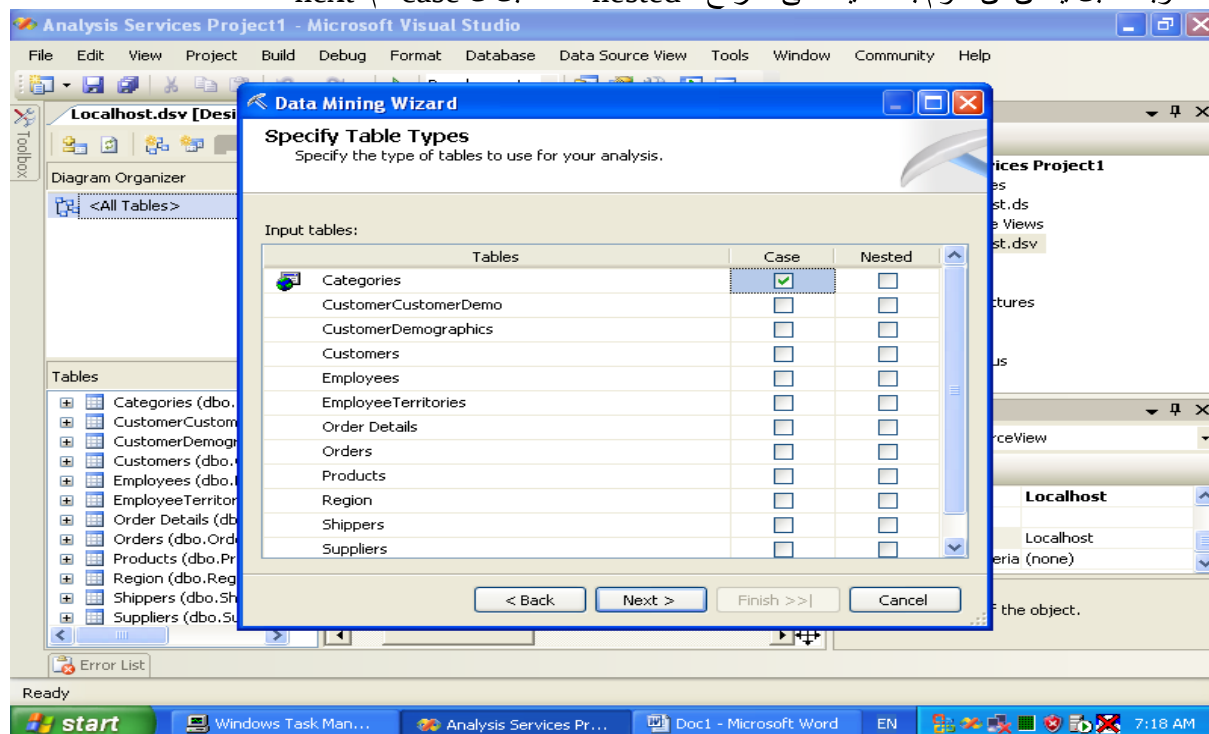
ملاحظه (يقصد بتقنية microsoft clustering عبارة عن تقنية تستخدم الخوارزميات لتجميع سجلات من مجموعة قواعد بيانات الي تجميع معين من النوع يسمى string ثم ليتم ذلك انقر على next)



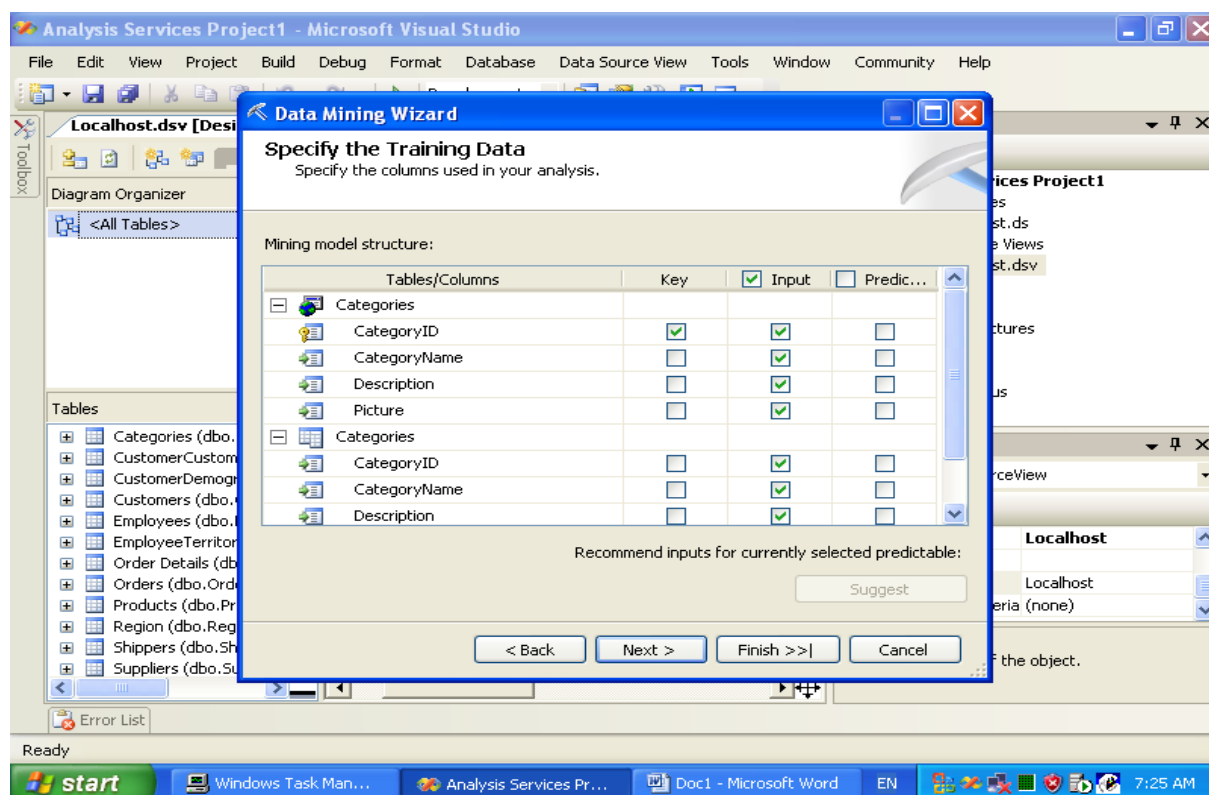
٥- يتم لنا عرض مربع حوار يحتوي على مجموعة من data source views يمكن ان نقوم بتحديد احد هذه المصادر ثم الضغط على زر next

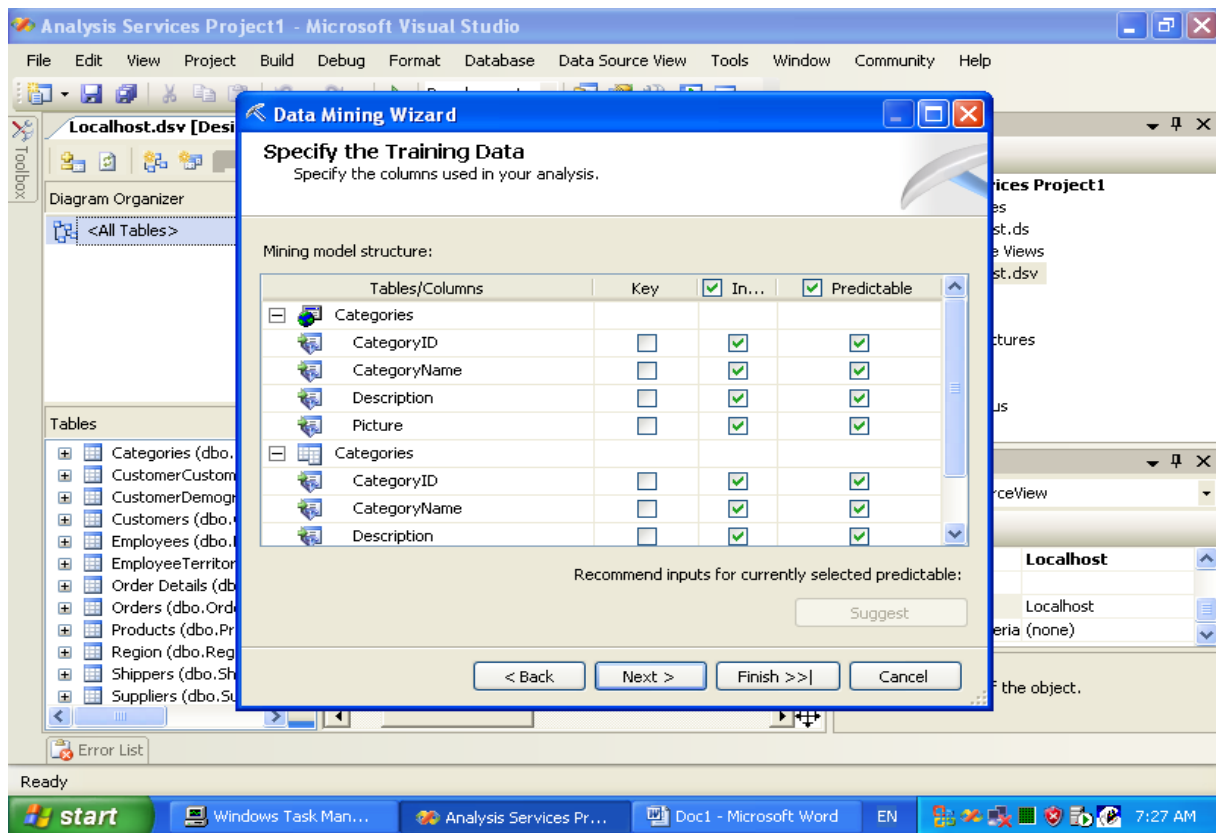


٦- يظهر لنا مربع حوار يحتوي على input tables ومن خلال هذا مربع الحوار نستطيع ان نقوم بالتحديد على الجداول التي نريد ان نستخدمها في عملية تحليل البيانات وذلك عن طريق تفعيل خيار الـ case وبالمقابل يمكن ان نقوم بالتحديد على الفراغ nested المقابل لـ case ثم next

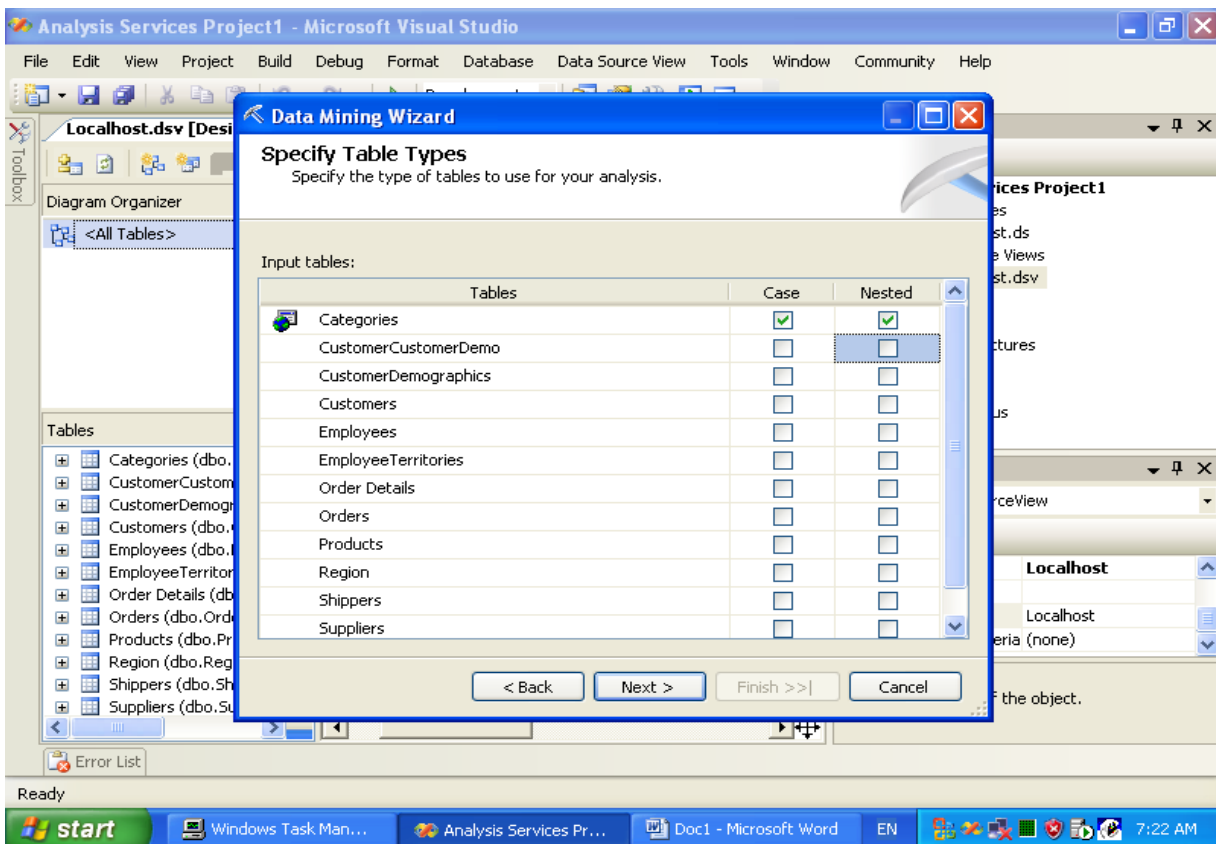


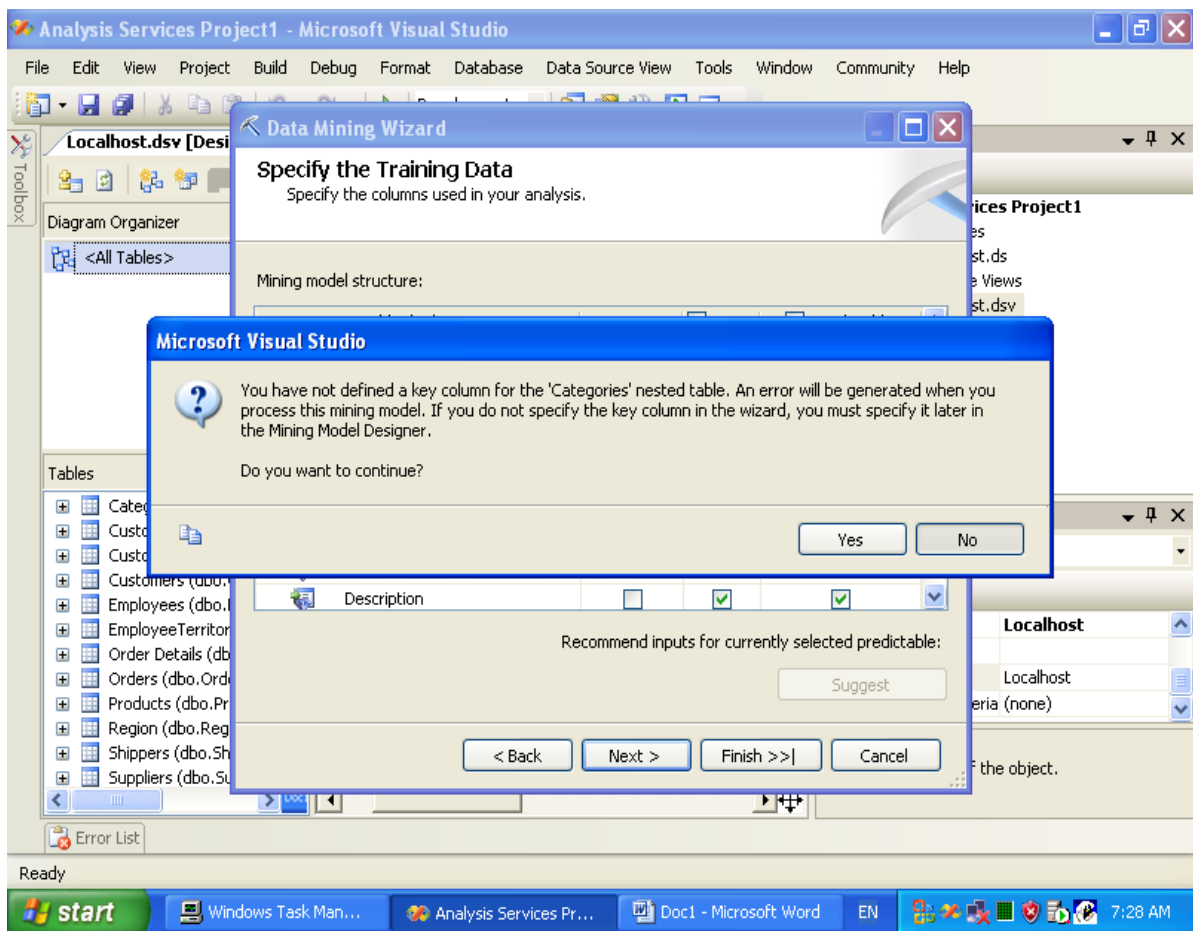
٧- نقوم بتحديد الاعمدة التي سوف تستخدم في عملية التحليل وذلك عن طريق التحديد على عمود الـ input واذا اردنا استخدام بعض الحقول لاستخدامها في تحليل تنبؤات في اتخاذ قرار في المستقبل فنقوم على التحديد على حقل predictable وبعد ذلك نقوم بالضغط على زر next ثم finish وهنا يظهر المطلوب





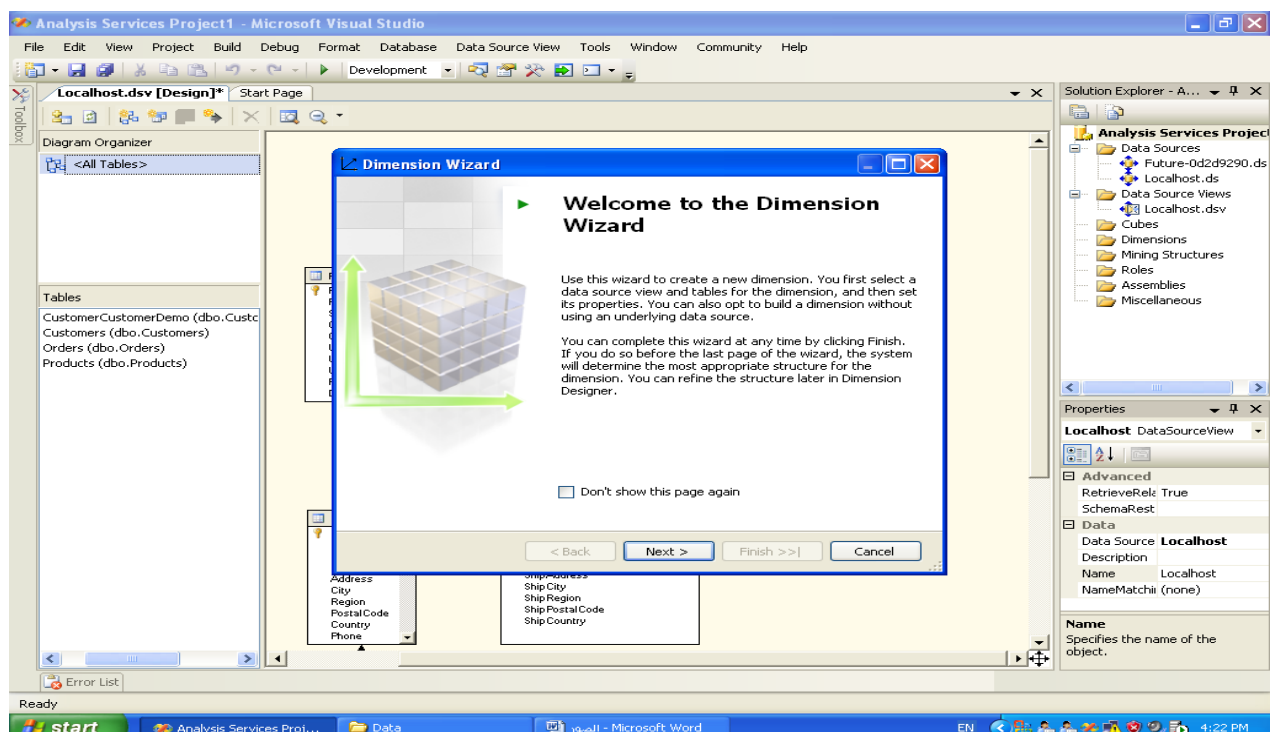
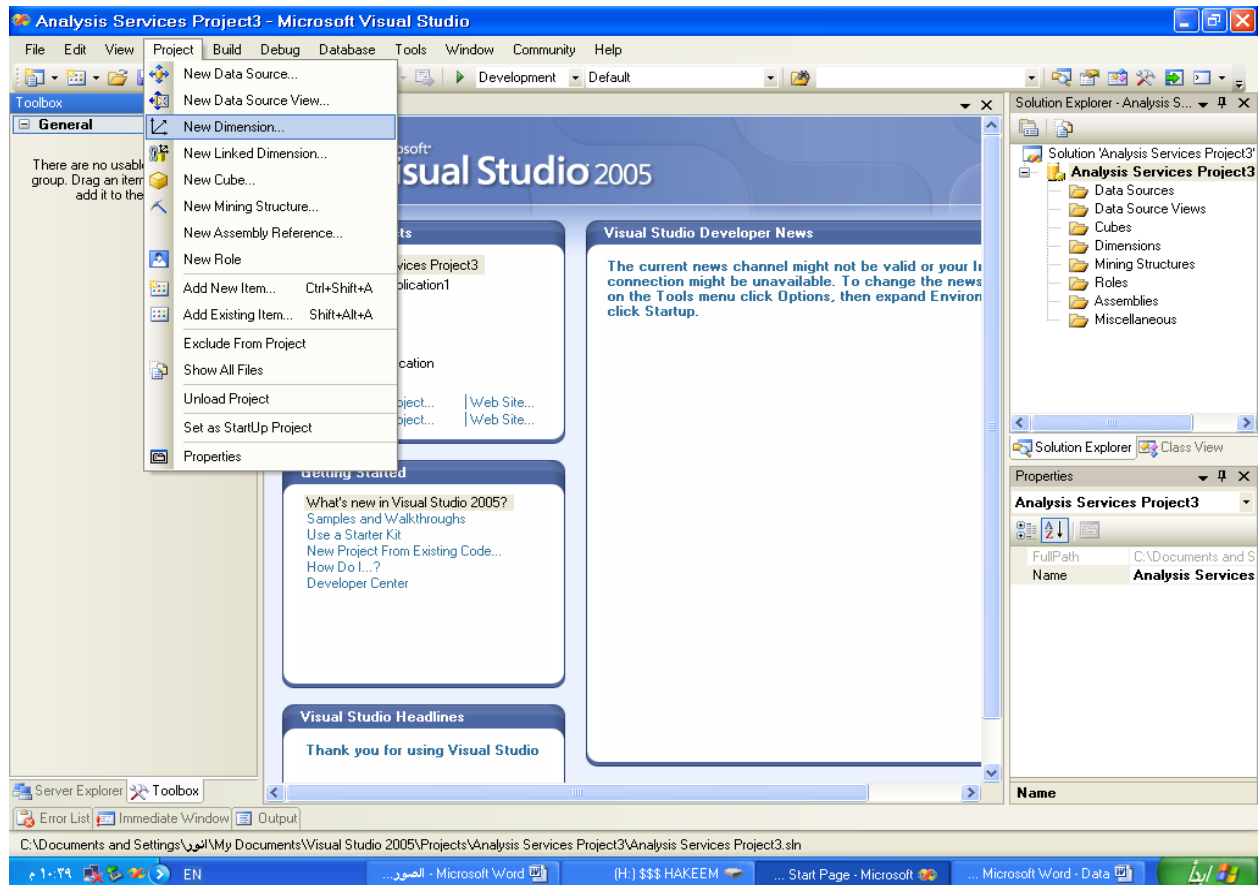
٨- من خلال مربع الحوار

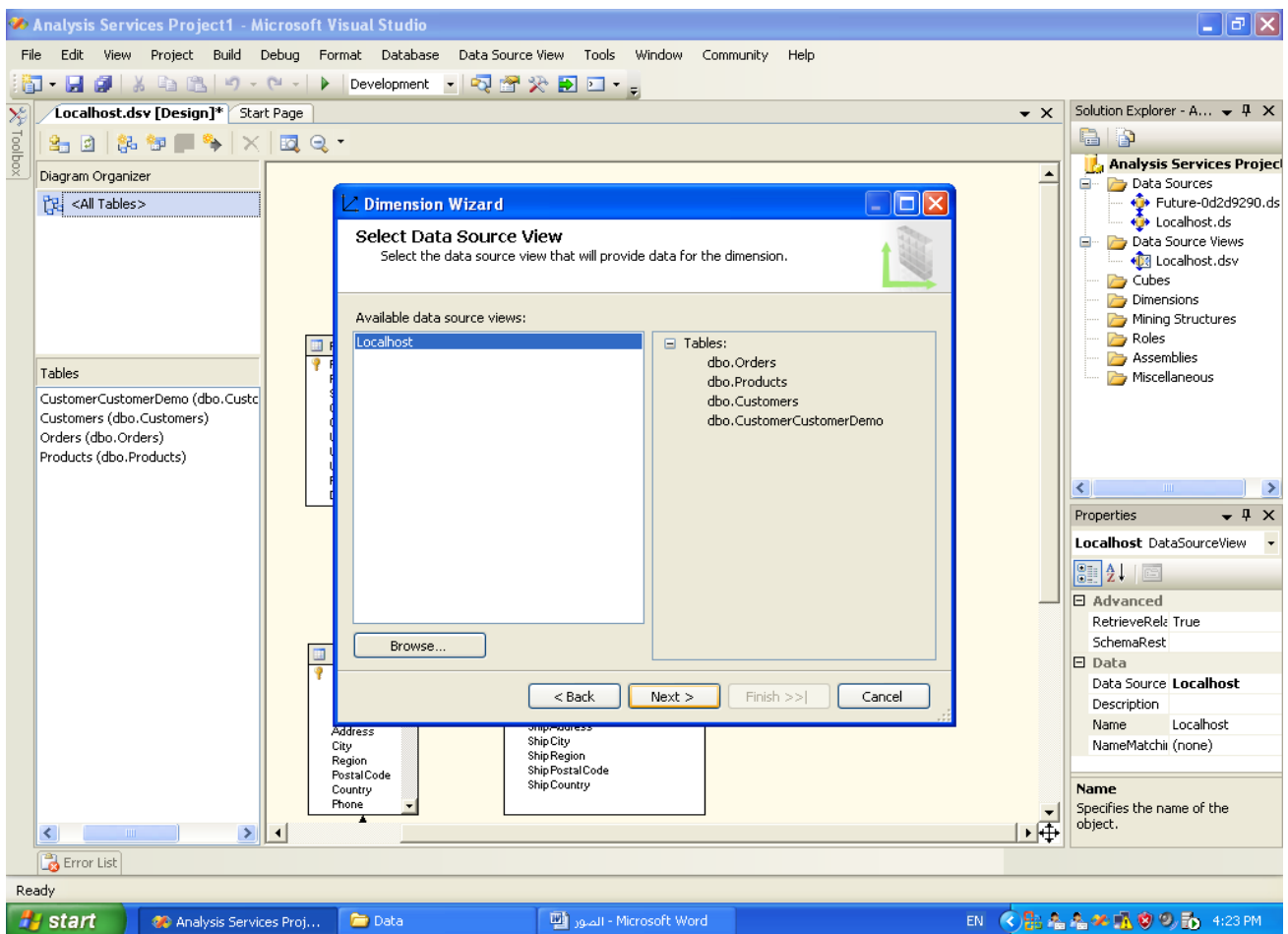
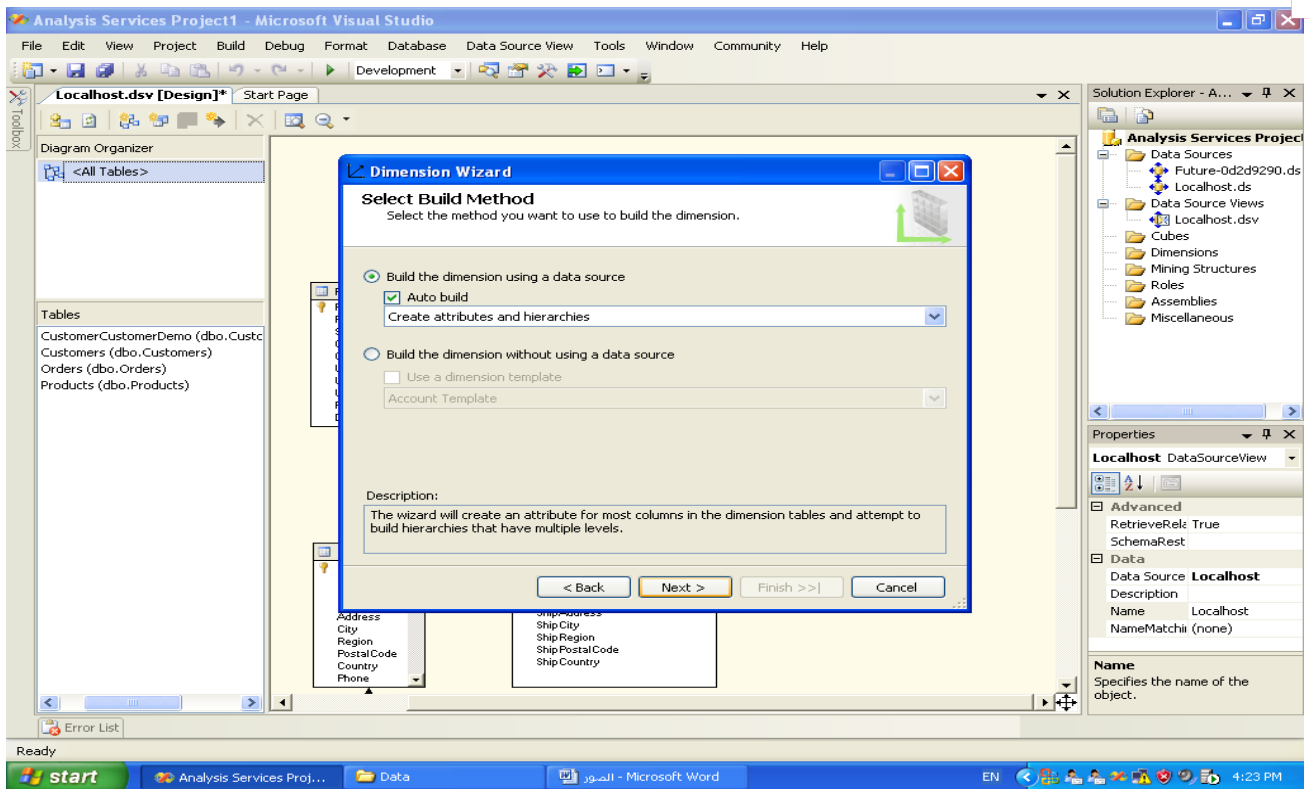


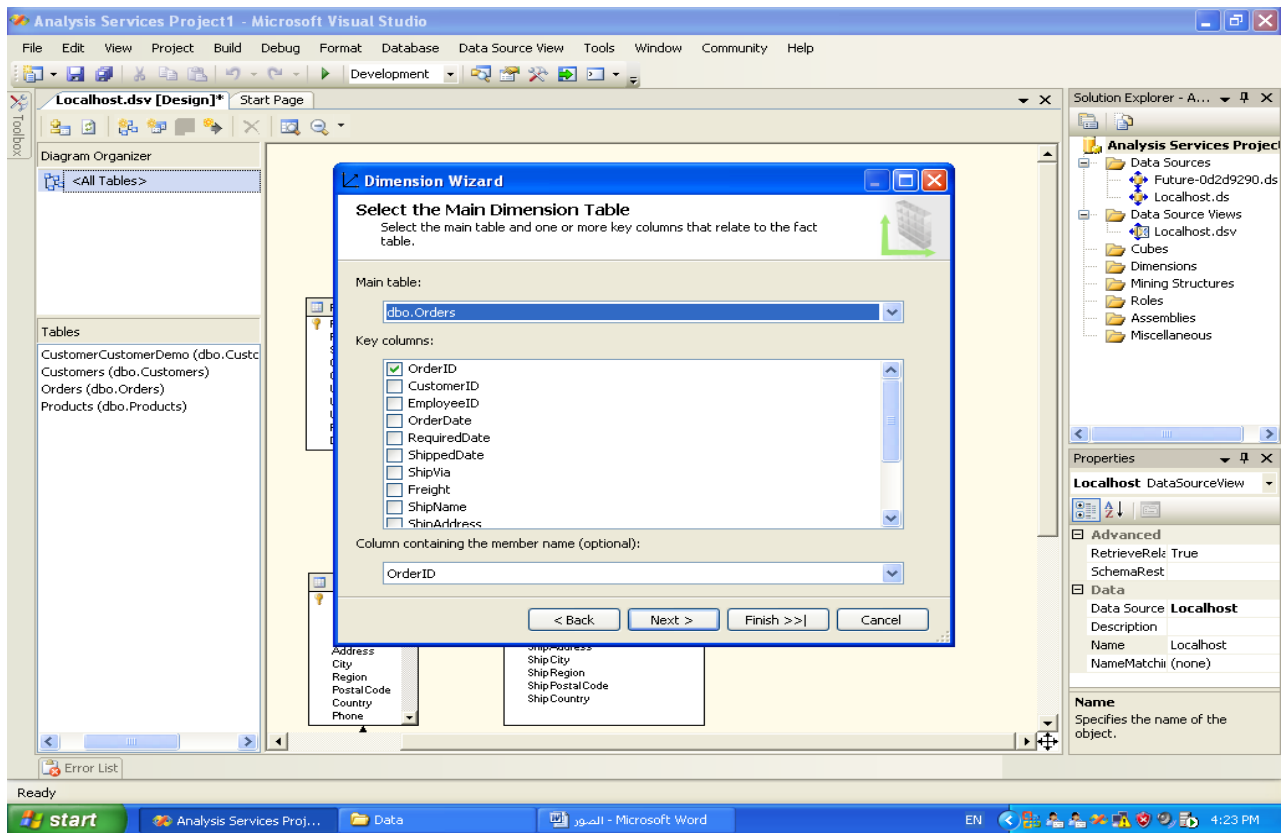
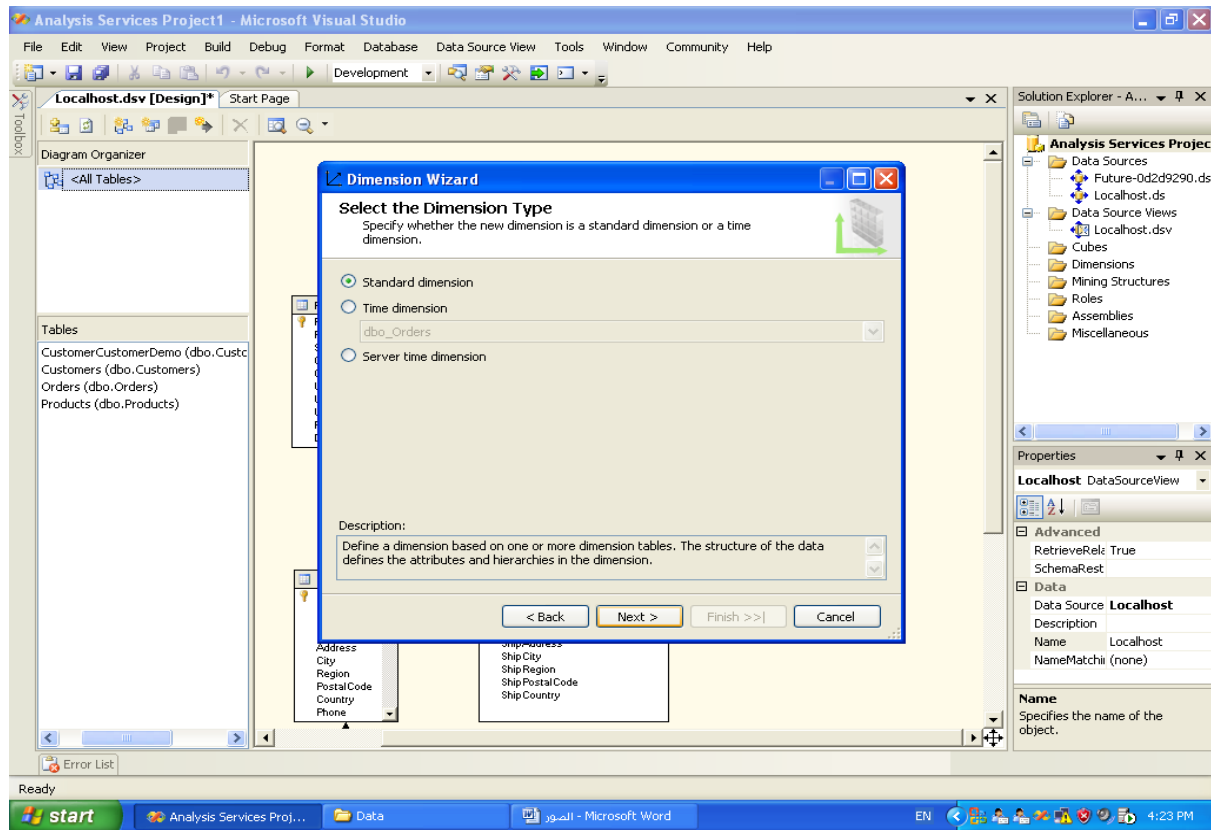


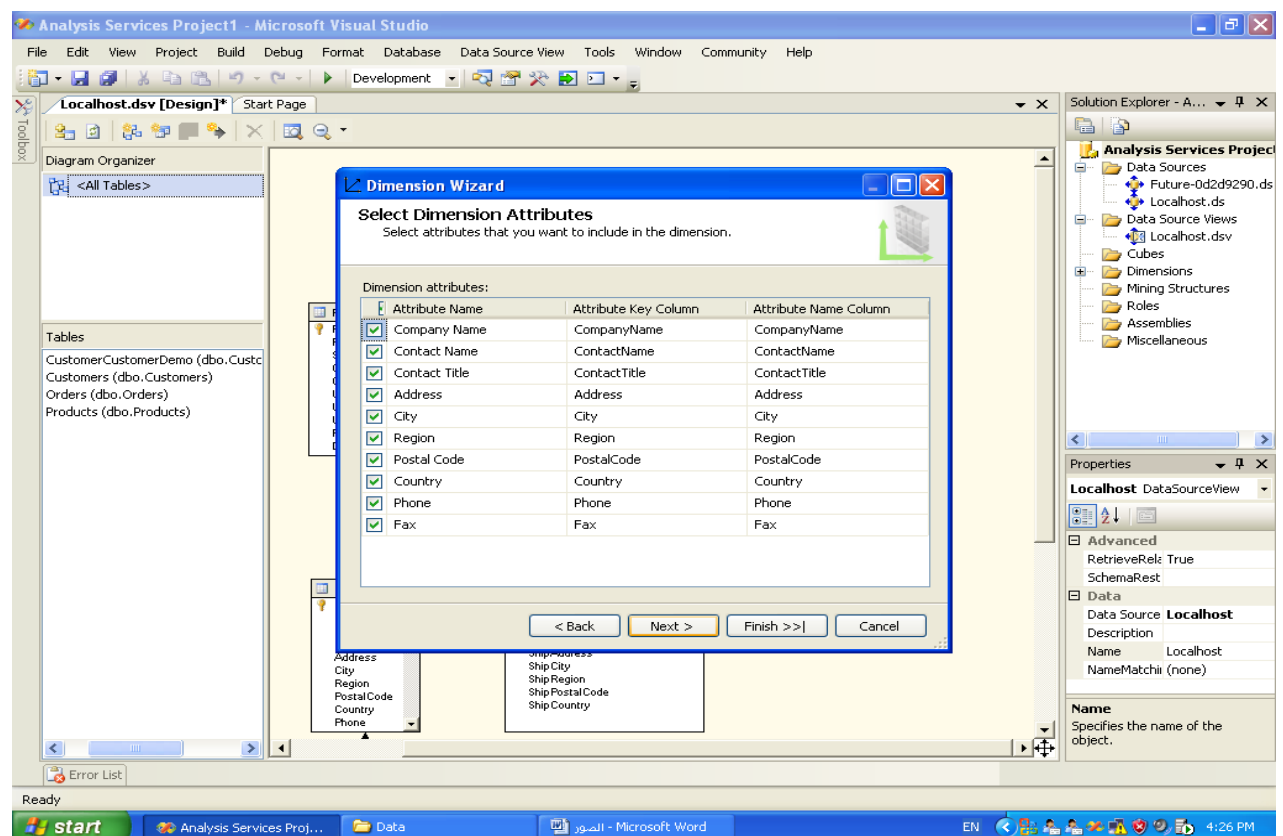
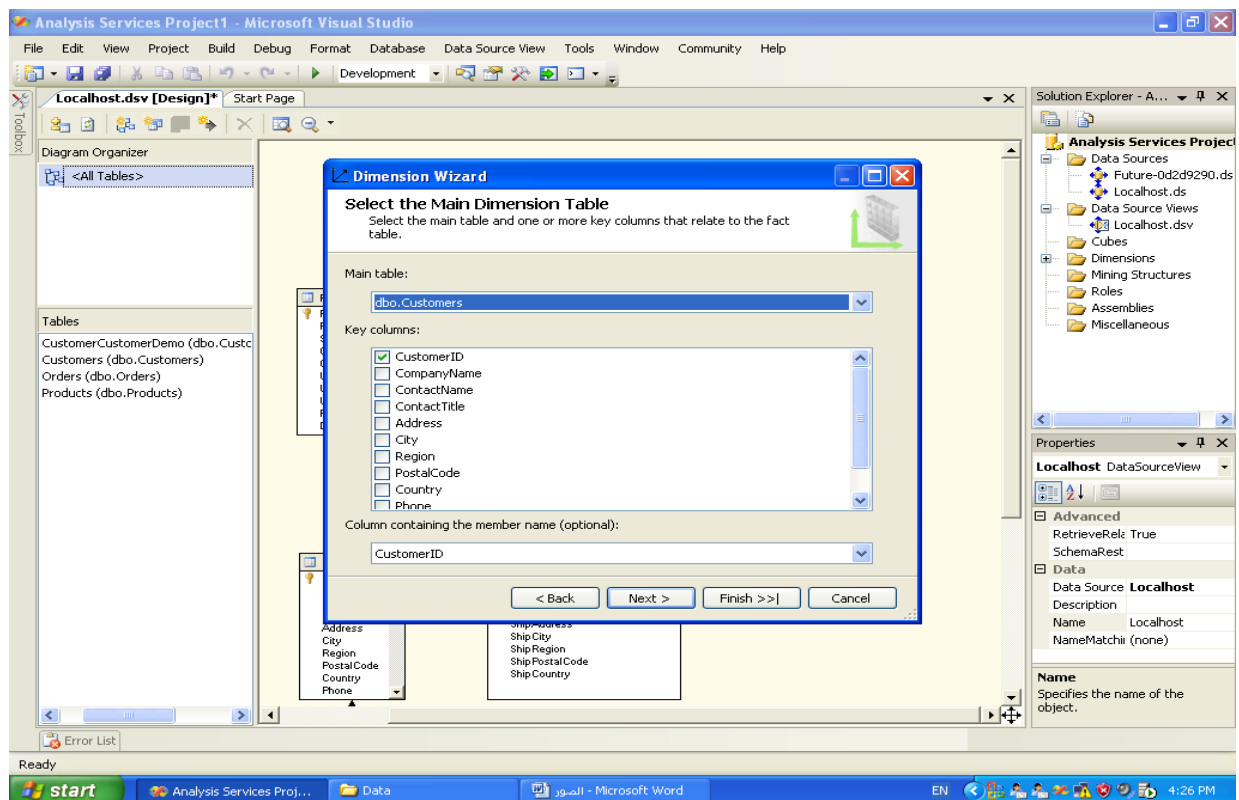
عمل New Dimension

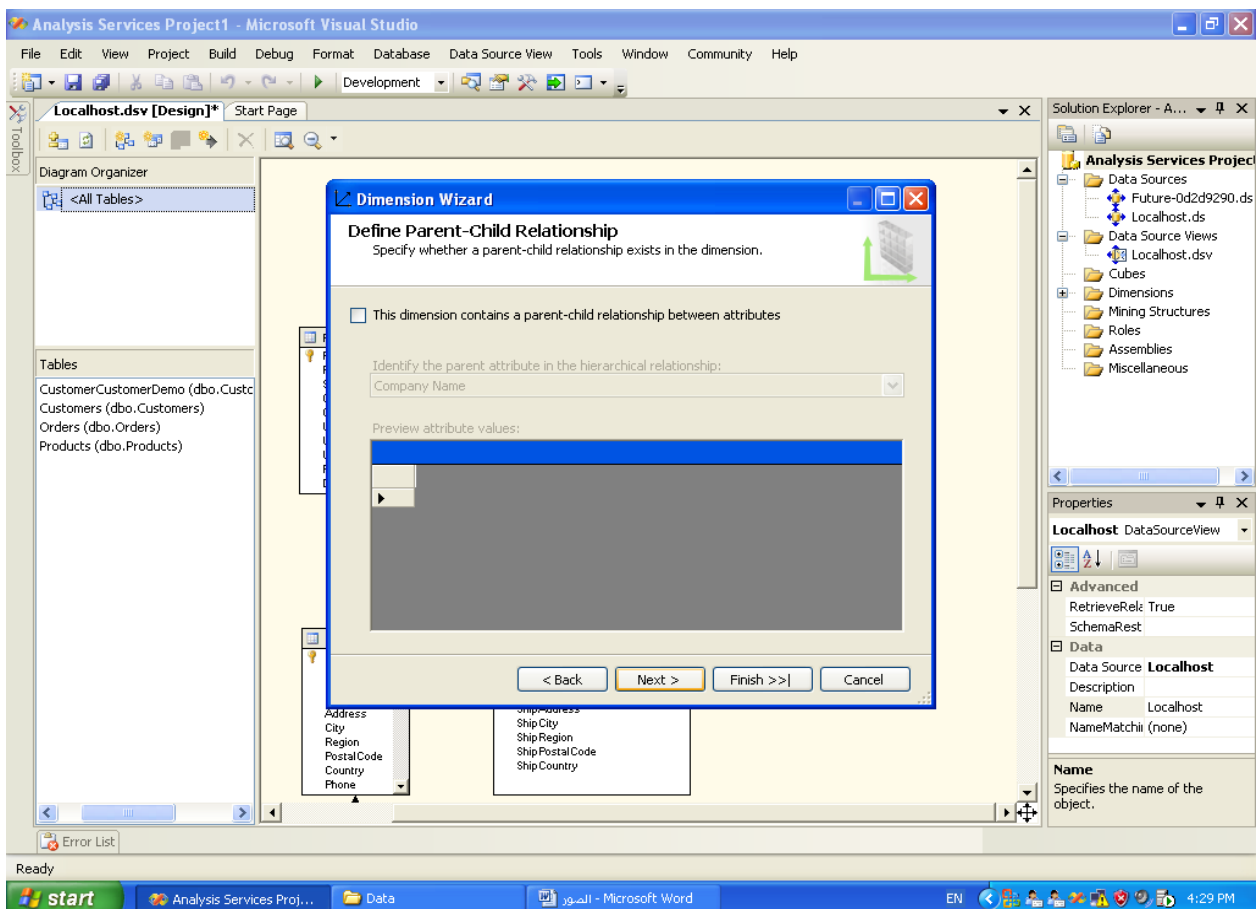
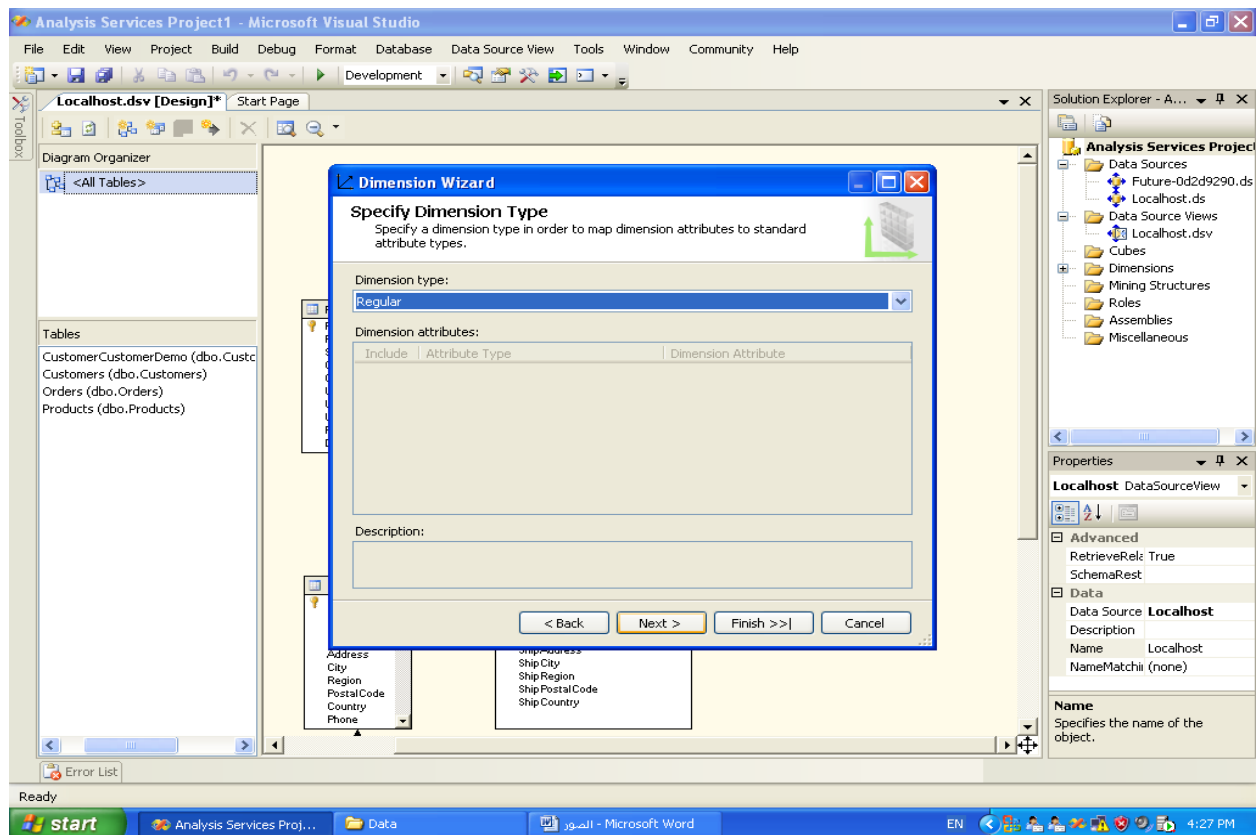
من قائمة project
نختار New Dimension











✓ This dimension contains a parent-child relationship between attributes

ثم نؤشر علامة صح

